An Overview of Machine Learning Methods for Assessing Movements in Housing Prices in Ireland

Barış Engin
University of Limerick
Limerick, Ireland
24213918@studentmail.ul.ie

Anas Mentak
University of Limerick
Limerick, Ireland
18033121@studentmail.ul.ie

Sebastian Alejandro Rodríguez Corona
University of Limerick
Limerick, Ireland
24036447@studentmail.ul.ie

Peter Debono
University of Limerick
Limerick, Ireland
24165239@studentmail.ul.ie

***Abstract*—With housing price increases being seen every year, policymakers need to take action to take targeted measures to address what has been defined as a housing crisis in Ireland. Machine learning algorithms can help policymakers to take targeted action where the most urgent interventions are needed. It can also assist other stakeholders to assess the overall health of the housing market. In this paper, multiple techniques are applied to a purposefully consolidated dataset consisting of various variables which affect housing prices. The results of the techniques show that tuned versions of boosted trees, particularly of Gradient Boosting and XGBoost, delivered the best results. This demonstrates that a novel approach towards the assessment of housing prices is possible and allows stakeholders to base decisions based on accurate forecasts.**

***Keywords*—machine learning, housing prices, housing market, regression, clustering, gradient boosting, boosting, bagging, XGBoost, stacking, stepwise regression, Ireland**

I. Introduction

The cost and affordability of housing is becoming a concern amongst EU citizens, particularly for young people. According to “The 2023 State of Housing in Europe” report, house prices in the EU increased by 47% from 2010 to 2022. [1]

The data in Ireland tells a similar story. According to the Irish Central Statistics Office, the price of an average dwelling for the 12 months leading up to January 2023 was €318,000. In 2012 the price was €205,476, representing an increase of 48%. [2] [3]

Furthermore, Ireland is currently experiencing a housing crisis. Savills, an international property adviser, undertook a study across eight developed countries and found that during the years 2015 to 2023, for every new unit of housing built in Ireland, 3.8 people were added to the population, the highest rate amongst the surveyed countries.[1]

Considering that there is the availability of data on the subject matter and keeping in mind the importance of the subject for policymaking, there is an opportunity for research to be undertaken on this. Machine learning is ideal for this task as past data can be used to help with informing future trends. The output of the models may be used by policymakers to determine which counties in Ireland are seeing the largest movements in house prices and take appropriate policy measures to address those increases. Through machine learning, the decisions which may be taken through the model would be well informed as the output of the machine learning model(s) would consider various real world data points to arrive at that output.

The question that this research undertakes to answer is the following:

Can machine learning be used in conjunction with macroeconomic factors, such as past housing prices, demographic trends, crime rates and financial indices, to assess movements in housing prices and understand the health of the housing market?

II. Related Work

Alshammari used real estate transaction data from Saudi Arabia, consisting of 13 columns and around 60,000 instances. The data was run through Random Forest (RF), Decision Tree (DT) and Linear Regression (LR) models and evaluated on Mean Absolute Error (MAE), Mean Square Error (MSE) and the coefficient of determination (R^2). The research concluded that RF performed better than the other methods to predict house prices. [4]

Kim, Lee, Lee and Hong merged two datasets to create one containing real estate transaction data and building information data in Seoul, South Korea. A total of 287,034 rows, with 16 variables were prepared for assessment by RF, neural networks (NN), inverse distance weighting, and kriging. Their research suggests that RF results were slightly better than those for NN, however it was suggested that NN may be more accurate if more data were available. [5]

Ho, Tang and Wong took a sample of around 40,000 housing transactions taking place in Tseung Kwan O, Hong Kong across 18 years and applied RF, Gradient Boosting Machine (GBM) and support vector machine (SVM) models to the dataset. They concluded that RF and GBM models had a superior performance compared to SVM. [6]

[1] A. Smyth and J. Ring, “Irish population growth exceeds new home delivery by almost 4 to 1,” Savills Ireland. Accessed: Oct. 16, 2024. [Online]. Available: https://www.savills.ie/blog/article/365469/ireland-articles/irish-population-growth-exceeds-new-home-delivery.aspx

Phan applied machine learning techniques on a data set of Melbourne housing transactions, consisting of more than 20,000 observations and 11 features. It was observed that the best performing model was one which used a combination forward stepwise regression and SVM, when compared to Principal Component Analysis (PCA), LR, NN and regression tree (RT) models. The study also observed model runtimes on R and found that RT and NN were the best performers. It also acknowledges that SVM is more difficult to interpret than regression trees. [7]

Lahmiri, Bekiros and Avdoulas found that boosting ensemble regression trees performed best when compared to Gaussian progress regression and support vector regression. A limitation of this study is that the dataset it uses, consisting of property transactions in a particular residential district in Taipei, Taiwan, only consists of 414 rows and 5 attributes. [8]

Kalehbasti, Nikolenko and Rezaei used six different models to arrive at the proposed rental price for properties listed on Airbnb in New York City. Based on R² and Mean Squared error, they concluded that the support vector regression (SVR) model outperformed models such as K-means clustering with ridge regression and neural networks. Interestingly, this study makes use of sentiment analysis on Airbnb reviews to improve the performance of the models. [9]

Barlybayev, Sankibayev, Niyazova and Akimebova scraped apartment listings on an online marketplace in Astana, Kazakhstan, and carried out an analysis on a dataset of 18,992 entries with 24 features. They used 12 regression models and a Bayesian neural network. The Bayesian neural network outperformed the regression models. From the regression models, the boosted trees and bagged trees displayed the best results. [10]

Yucebas, Yalpir, Genc and Dogan take a different approach. They use a hybrid model combining X-means clustering and classification regression trees and compare it against a 'direct capitalisation', which attempts to estimate property values based on rental income. The model provided superior results over the direct capitalisation method. [11]

Park and Bae use a variety of models to predict housing prices of properties in Fairfax County, Virginia, USA. Their research, carried out on a dataset of around 5,300 records, concluded that the RIPPER model was the best model when compared to other models such as Naive Bayes, C4.5 and AdaBoost. [12]

Zhang uses what is known as the Spearman correlation coefficient to determine which factors most affect housing prices, and uses a multiple linear regression model to forecast housing prices from a housing price dataset in Boston, USA. Zhang concludes that multiple linear regression can be used to predict housing prices, however it is noted that there is room for improvement. [13]

This study aims to determine whether housing prices can be predicted by considering not only the prices of previous properties but also the following data: stock market prices, crime statistics, migration, unemployment rates, and the supply of new properties.

In relation to stock prices, multiple studies have found that there is a long-term positive relationship between stock market prices and housing prices. [14] [15] [16]

The negative effects of crime would suggest that there is a negative correlation between crime rates and housing prices. In fact, Zhang observed this as one of the negative correlational factors. [13]. Ceccato and Wilhelmsson observed this when studying data about apartment sales in Stockholm, Sweden. Housing prices were affected regardless of the type of crime committed. [17] Related to this, Barlybayev's study also contains data about individual security features of each property sale and found that the presence of security features correlated positively with housing prices. [10]

Conversely, the basic principle of demand and supply would suggest that there is a positive correlation between migration and housing prices. In fact, Sanchis-Guarner studied the impact of this across Spanish provinces between 2000 and 2012 and found that a 1% increase in the immigration rate increased average house prices by 3.3%, mainly due to new immigrants but also due to relocations by the native population. [18]

Lower rates of unemployment were seen to increase housing prices over the long term in a study performed by Gan, Wang and Zhang on data from cities in Texas. [19]

III. Methodology

A. An Overview of the KDD Process

The research was conducted using the Knowledge Discovery in Databases (KDD) process, which was established to find patterns and assist with the interpretation of large datasets. The term was coined in the late 1980's and the process was defined by Fayyad, Piatetsky-Shapiro and Smyth in 1996 in their seminal work, "From Data Mining to Knowledge Discovery in Databases". [20]

The process involves nine steps, and these will be documented throughout the course of the paper. The authors of the KDD process have stipulated that the process can involve loops between any of the steps.

B. First Steps

The first step involved understanding the application domain, prior knowledge, and goals of the research. In this case, an understanding of the relationship between the predictors of housing prices and the response of said prices was developed through a basic understanding of the factors affecting house prices and also through review of existing literature, on the relationship of factors such as crime and stock prices. The goal is to apply this understanding and gained knowledge to the research question.

The second step was to create the target dataset. To answer the research question, it was established that the following types of data were needed: housing transactions, crime, immigration, unemployment, new housing builds, and stock market data. The data for the first five parameters was readily available from the Central Statistics Office of Ireland. [21] [22] [23] [24] [25] [26] [27] The stock market data was retrieved from Yahoo Finance through Python. [28]

C. Data Cleaning, Preprocessing, Data Reduction and Projection

The third step of the KDD process involved data cleaning and preprocessing. Although the data was easily retrievable, various steps were necessary to ensure optimal machine readability.

Prior to performing the cleaning and preprocessing the data had to be understood. Each dataset was loaded in a separate dataframe by means of the Pandas package in Python. Then, the datasets were checked to identify whether they were recognised as objects or integers, along with a manual review of the whole datasets to see whether there were any errors missing fields.

Normalisation had to be performed on the datasets to ensure machine readability. For each dataframe, each column name was converted to lowercase. On the Housing dataset, columns were stripped of extra blank spaces. The datatypes in this table were being seen as objects therefore some adjustments had to be made. The date and time column was converted to a date format, and the month and year were extracted separately. The comma present in the price values was removed; the datatype converted to float and rounded to 2 decimal places.

Some property descriptions (new or second hand) were written in Irish and were replaced with the English translation. Features which had Null price, county, date or property descriptions were deleted to ensure consistency. Finally, a check was performed on the counties to ensure that all 26 counties were present in the housing data.

For the Finance and Crime data, dates were converted from an object to a date, and month and year were extracted to new columns. For Finance, Null values were filled in with the closing price of the previous day, whilst for Crime, features which had Null values for amounts and county were deleted. For Migration and Unemployment, any values which had Null values for county and value were deleted.

On the new builds data, the column for the quarter was split into Year and Quarter. Quarters were mapped according to months from 1 to 12. Then, each row was iterated over, using the quarter to month mapping, and the number of new builds being distributed evenly across the three months. The county names in this data had to be amended to match the other dataframes - Dun Laoghaire Rathdown, South Dublin and Fingal were renamed to Dublin, and any text which was not the name of the county (e.g. City/Council) was removed.

The fourth step of the KDD process involved data reduction and projection.

From the Housing dataset, columns such as street address and information about VAT exclusivity were removed. Each county name and each description of property had to be ‘mapped’ against a number to ensure optimal readability with any machine learning model used.

The initial Unemployment dataset contained columns which were out of scope of the analysis, such as marital status, age group, and sex. This data was removed. Furthermore, the dataset only contained values for 2011, 2016 and 2022. To fill in the missing data, all data was grouped by county, then, missing values were filled in by following the unemployment trend of Ireland from 2003 to 2024. This way, the missing years were approximated, and the accurate rates were kept.

A similar process followed for the Migration dataset. Rows containing census periods from 1951 to 2002 were removed. Since the census periods covered five years, these had to be expanded to cover five columns. For example, years 2011-2016 became different rows containing each of the separate years. Similarly to the Unemployment dataset, missing values for the newly created years, 2023 and 2024 were estimated by following the migration trend of Ireland from 2003 to 2024 and mapped it on to every county.

For the Crime dataset, again, out of scope columns were removed. Every crime was mapped to a less specific crime type. Additionally, the types of crimes were added together to capture the total crimes committed per month, per county. Since the data rows included the Garda division as a location, these had to be mapped to counties.

The Crime, Migration and Availability datasets were further reduced to capture only instances from 2019 onwards, to ensure they match the Housing data.

The Availability data was transformed to contain a row for every combination of month, year, county, and number of new houses. Zeros were included in parts where no data was present for that combination. A similar process was undertaken on the Migration and the Unemployment data however, since the migration data was based on annual movements, this was divided by 12 to be distributed evenly across the 12 months in a year.

All the transformed datasets (excluding Financial data) were merged into one. A new column called ‘period’ was added to give a specific numeric value to each month from January 2019 onwards. At this point, the 2024 data was removed as the data for the year was incomplete at the time of carrying out the study. Furthermore, financial data was not included as this would be added later, after the data was organised into clusters.

The cleaning and transformation process for the finance data involved some steps. The finance data retrieved only included data from August 2019 onwards therefore a mean of the remainder of the 2019 data was taken and applied to rows from January to July 2019. The 2024 data was also trimmed out. A ‘period’ was also assigned to the financial data. At this stage, the training data consisted of 293,926 rows and 12 columns. A sample of the data is shown in Table I below:

TABLE I SAMPLE OF AGGREGATED DATA (PRIOR TO CLUSTERING AND INCLUSION OF INDICES)

	year	month	period	county	county_id	type_of_property*	type_id	price	total_crimes	new_houses	migration_thousands	unemployment_rate
0	2019	1	1	Donegal	5	Second-Hand Dwelling house /Apartment*	2	47400.00	2458	28.0	0.41	1.01
1	2019	1	1	Kilkenny	10		2	60671.00	1227	18.0	0.34	0.72
2	2019	1	1	Dublin	6		2	30000.00	47638	487.0	0.61	0.62
3	2019	1	1	Dublin	6		2	30000.00	47638	487.0	0.61	0.62
4	2019	1	1	Waterford	23		2	130000.00	2585	25.0	0.99	0.8

*Note: all 5 rows contained the same type_of_property, hence for the sake of illustration and ease of formatting they appear as merged in this table.

At this stage it was also decided that models were also to be run with on a variation of the dataframe reflecting the mean price per county per month.

Furthermore, it was decided that the models were to be run on another version of the dataset, with outliers removed. After reviewing the data through a box plot, it was decided that prices below €150,000 and above €4,000,000 were to be removed from this alternative dataset.

A correlation matrix was plotted to confirm if any multicollinearity existed between any of the predictor variables. After this, a K-means clustering model was run on the dataset. This will be explained in more detail in the sixth and seventh step of the KDD process, namely exploratory analysis and data mining. The four clusters created by the K-means clustering were saved into separate dataframes for individual analysis later.

After the clustering process was completed, the separate clusters were again merged into one dataset, and the cleaned and transformed finance data was incorporated in the dataset.

Therefore, the following (Table II) were the features of the data after clustering and before further implementation was carried out (17 features):

TABLE II FEATURES AND DATATYPES AFTER CLUSTERING AND FINANCIAL INDICES

period	int32
county_id	int64
type_id	int64
price	float64
total_crimes	int64
new_houses	float64
migration_thousands	float64
unemployment_rate	float64
cluster_label	int32
tsx	float64
ftse	float64
nikkei	float64
vwo	float64
kosdaq	float64
bcom	float64
glab.l	float64
vix	float64

D. Goals, Identification of Machine Learning Methods, and Exploratory Data Analysis

The fifth step of the KDD process involved establishing its goals. For the analysis undertaken, two goals were identified. The first goal of machine learning in this research was to apply clustering. This was necessary to better discover the relationships between the other variables, not just price, especially the economic variables being used may have other distinct links which may not be immediately obvious.

The second goal was regression, which maps one or more predictor variables to produce an output for a response variable. The data consisted of many predictor variables (past prices, migration, crime, financial data, unemployment, availability of housing stock) and the final output variable was a house price.

The sixth step of the KDD process involved exploratory analysis and the selection of machine learning methods. Firstly, the machine learning methods shall be identified as follows:

K-means clustering - Clustering would allow particular points in the datasets to be grouped distinctly. For K-means clustering to be performed, the number of clusters must be specified. The K-means algorithm will select the data point and assign it to a cluster depending on its proximity to other data points. [29] The rationale behind using this is to identify whether any trends can be identified from the data.

Stepwise regression - In regression one takes predictor variables to predict the output of a response variable. In cases where more than one predictor variable is present in the data, it is a good idea to identify the variables most relevant to the computation of results. Stepwise regression was identified as a method which would be easier on resources than best subset selection. [29] Backwards stepwise regression was carried out on the data, which involves starting to look at the data with all input variables and eliminating the less relevant ones until the most relevant variables are determined.

Stacking regressor – A stacking regressor is an ‘ensemble’ machine learning method which combines the output of multiple models as an input and combines these to improve the performance of the overall model. [30] The premise of this model, being the fact that it combined multiple machine learning techniques made it worth investigating.

Bagging – This is a decision tree method of machine learning which is based on the bootstrap concept, which involves getting unique datasets, being the same size as the original dataset, but using sampling with replacement from the original dataset. Models are then trained on each gathered dataset and the error is calculated according to the average of each model. [29] As the precursor to the Random Forest technique, it was seen as a good starting point.

Random Forest (RF) - RF is a tree-based method, which involves the segmentation of the input variables into different groups. It is a better version of bagging. RF averages the results of training of multiple models from random subsets of the original datasets, not from the whole datasets, in order to reduce variance. [29] In some of the studies, the RF method provided the most accurate results, therefore it was worth investigating the performance of this model. [4] [5] [6]

Boosting – This is also a tree-based method. However, where bagging methods use different subsets of the original data to train the model, boosting develops the model sequentially, as each decision tree is modelled on the basis of the previous tree. [29] Boosting was also observed to be successful in multiple studies and was also worth investigating. [8] [10] Two implementations of Boosting were used namely the Gradient Boosting regressor and XGBoost.

Support Vector Machines (SVM) – This is a method which utilises what is known as a hyperplane, which in simple terms is a line which separates classes of data. The SVM aims to find the most optimal hyperplane automatically, it being the line with the maximum distance between the different classes. SVM is better than the support vector classifier as SVM can create non-linear boundaries between the classes. [29]

E. Exploratory Data Analysis and Data Mining

The seventh and eight step of the KDD process involve data mining and the interpretation of mined patterns respectively. The processes described below will involve looping between the steps just described, as well as with and the data exploration part discussed in the sixth step.

Before regression could take place, the data was organised into clusters through K-means clustering. First, a separate dataframe was created for the features to be clustered, without the house price. In this case, clustering was done to observe whether any other groupings existed between the predictor variables, apart from the price.

The data was scaled to ensure standardisation and avoid larger variables, such as crimes, from dominating the clustering process. Then, a K-means model, and it was specified that 4 clusters were to be created. Then, the dimensionality of the data was reduced to 2 components by means of Principal Component Analysis (PCA). The components were then plot against each other to visualise the relationships between each other.

In defining a suitable approach, different data distributions were experimented on. First, machine learning was conducted on individual house prices; 293k observations over approximately a 4.5-year period. The results exhibited excessive unexplainable variance. This dataset included all prices, with a minimum price of €30,000 and a maximum price of €200,000,000. Large dispersion in distribution around the mean affects training greatly. However, with respect to assessing the health of such a small and compact market with low supply, every property price matters. Nonetheless, the models were trained on the trimmed dataset where outliers were trimmed to limit the data within the range of €150,000 and €4,000,000. Results were unconvincing, and high variance persisted which led the approach towards training on mean prices rather than individual observations. Interestingly, mean price dataset outperformed trimmed mean prices in training.

Consequently, significant learning opportunities were observed and pursued using clustered mean property prices dataset. It is crucial to recall at this stage that this research is intended to inform policymakers on the health of the housing market from a macro-economic point a view. Therefore, the choice of methodology must align with this objective.

To examine the effects of movements in international markets on the price levels in the Irish housing market, the clustered data was merged with the 8 indices defined in prior sections. Data was then split into 4 different datasets as per the number of clusters. Initial learning was conducted on the cluster level, followed by learning on all clusters, before defining statistically significant parameters using a backwards stepwise regression and proceeding to training on optimal features.

The learning process was completed using 6 different models, which were carefully selected given the complexity of the observed relationships in our data. The following are the models used and their parameters:

1. Boosting

a. GradientBoostingRegressor

```
gradient_boosting_regression(X_train, y_train, X_test, y_test, n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42)
```

b. XGBRegressor

```
xgboost_regression(X_train, y_train, X_test, y_test, n_estimators=100, learning_rate=0.1, random_state=42)
```

2. 2. Decision Trees

a. Random Forest

```
RFR(X_train, y_train, X_test, y_test, n_estimators=100, random_state=42)
```

b. BaggingRegressor

```
bagging_regressor(X_train, y_train, X_test, y_test, estimator=None, n_estimators=10, random_state=42)
```

The estimator parameter set to none indicates DecisionTreeRegressor as the default base estimator.

3. Stacking Regressor

Outputs from three base learners were combined using a linear regression to estimate the stacked models' output. The three models in question are:

```
def stacking_regressor(X_train, y_train,
X_test, y_test):
    # Define base learners
    base_learners = [
        ('rf',
RandomForestRegressor(n_estimators=100,
random_state=42)),
        ('gb',
GradientBoostingRegressor(n_estimators=1
00, learning_rate=0.1, random_state=42)),
        ('svr', SVR(kernel='linear'))
    ]
```

The choice of base learners followed results obtained from GBR and RFR implementations, supported by an SVR with a linear kernel.

4. SVR; Kernel = RBF

In addition to using a linear kernel within an SVR model in the stacking regressor, another application of this model was used with an RBF kernel to assess the model's suitability for this problem. A scaler object was used within a pipeline to prepare data for training.

The training datasets are structured as follows:

TABLE III. STRUCTURE OF TRAINING DATASETS

Dataset	Observations	Mean	Standard Deviation
Clus. 0	2874	242720.53	88902.55
Clus. 1	66	664873.77	282569.02
Clus. 2	40	647025.57	335694.26
Clus. 3	98	916950.06	347276.9
All	3078	278493.38	182872.01

IV. EVALUATION

A. Evaluation Metrics

R Squared (R²)

R² shows the extent to which the variance in the target variable can be explained by the predictor variables.

The formula is the following:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}_i\right)^2}$$

RSS (Residual Sum of Squares) is the sum of squared residuals, showing the prediction error of a model, whilst TSS is the Total Sum of Squares.

Therefore, a value of 1 indicates that all variance in the target variable can be explained by the predictor variables, whilst 0 shows that none of the variance can be explained by the target variables.

A higher R² generally represents a better model. However, the R² does not account for any bias, therefore it is ideally used in conjunction with other evaluation methods to assess performance.[2]

Mean Squared Error (MSE)

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}i\right)^2$$

The MSE is related to the RSS, however, the MSE illustrates the average error, rather than the total error. The mean is used in order to enable the measure to be independent of the size of the dataset. The MSE is not shown in the original units presented in the dataset, therefore, interpretation is more challenging. Furthermore, since the value is squared, large errors, potentially arising from outliers, may have a heavy effect on the MSE. [2]

Root Mean Squared Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}$$

RMSE is the square root of MSE. This has the effect of bringing the value back to the original unit of measurement presented in the dataset, making it easier to read.[2]

Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}i\right|$$

MAE is similar to MSE, however instead of taking the sum of residual squares, the sum of absolute values is used. The effect of this is that similar to RMSE, the value can be interpreted at the same unit of measurement presented in the dataset. Unlike the MSE, the lack of squaring means that errors are treated equally, therefore making it stronger against outliers. [2]

[2] E. Lewinson, "A Comprehensive Overview of Regression Evaluation Metrics," NVIDIA Technical Blog. Accessed: Oct. 28, 2024. [Online]. Available: https://developer.nvidia.com/blog/a-comprehensive-overview-of-regression-evaluation-metrics/

B. *Results*

This section is fully focused on the eighth step of the KDD process namely the interpretation of mined patterns.

On the cluster level, cluster 0 showed the best learning results. The dataset's features explain between 60% and 66% of movements in the mean house prices during the 5-year period of training.

This cluster had the highest number of observations out of all clusters. A simple interpretation is that prices of houses in these areas that fall in the range of the observed mean +/- deviations can be explained by the learning features defined in this research. However, the learning accuracy is quite low for such interpretations. More implementations and evaluations were conducted to ensure thorough analysis.

TABLE IV CLUSTER 0

Model	R^2	RMSE	MAE
GBR	0.656	53296.59	33757.39
XGB	0.645	54112.61	34227.83
RF	0.65	53901.05	32925.47
Bagging	0.61	56416.84	34263.5
Stacking	0.66	52592.63	32813.59
SVM	0.02	89905.8	66262.3

The remaining three clusters showed insignificant results. This could be mainly attributed to the small number of observations in each cluster with 66 observations in cluster 1, 40 observations in cluster 2, and 98 in cluster 3. The latter has the highest mean prices, followed by cluster 2 and cluster 1.

TABLE V CLUSTER 1

Model	R2	RMSE	MAE
GBR	-1.096	259387.56	172135.21
XGB	-0.43	214175.29	123975.36
RF	0.24	156299.02	106825.68
Bagging	-0.1	188227	117446.8
Stacking	0.55	120060.41	95174.51
SVM	-0.03	182558.2	113218.7

TABLE VI CLUSTER 2

Model	R2	RMSE	MAE
GBR	-1.03	663249.98	434364.28
XGB	-1.31	707347.61	481835.09
RF	-0.59	587250.33	408305.8
Bagging	-0.7	608624	435415.2
Stacking	-0.27	525784.63	374796.17
SVM	-0.34	539853.14	329089.4

TABLE VII CLUSTER 3

Model	R2	RMSE	MAE
GBR	-0.588	406287.75	351860.09
XGboost	-2.21	577868.29	463855.20
RF	-0.54	400469.2	335577.2
Bagging	-0.77	429307.79	362684.2
Stacking	-0.24	358242	265034.86
SVM	-0.32	370786.5	260818.1

Following these results, learning was then conducted on the concatenation of all clusters. Higher performance was observed, with an average of 0.1 improvement in each model's R^2 measure compared to training on cluster 0, except for SVM. However, the Mean Average Error for the same individual cluster was lower than when combined with the remaining clusters. This indicates the presence of noise, for which attempt to reduce was made by extracting the optimal features for learning by testing for their effect using a Backwards Stepwise Regression.

TABLE VIII ALL CLUSTERS

Model	R2	RMSE	MAE
GBR	0.74	84929.43	45158.27
XGboost	0.68	94723.65	45312.86
RF	0.72	88450.3	43709.2
Bagging	0.69	93309.2	45826.8
Stacking	0.74	85343.24	43053.27
SVM	-0.02	169917	93187.3

The stepwise regression removed seven statistically insignificant parameters. The remaining features are Date, County, Type, Total Crimes, New Builds, Unemployment Rate, Cluster Labels, FTSE 100 Index, VWO Emerging Markets Index.

Results obtained indicate better training performance when reducing the dataset to statistically significant parameters only.

Both XGBoost and GBR were selected for cross-validation for parameter tuning. Although XGboost had a lower R^2 for the optimised dataset, it showed consistently meaningful results that motivated the hypothesis that better performance can be obtained by tuning learning parameters.

TABLE IX BEST FEATURES

Model	R2	RMSE	MAE
GBR	0.744	84947.03	44771.24
XGboost	0.62	102742.24	47379.83
RF	0.7	92384.7	45619.2
Bagging	0.66	98134.6	48337.2
Stacking	0.73	87150.92	44095.58
SVM	-0.17	169448	92740.7

Grid search k-fold cross-validation with k = 5 yields the following parameters for each of the specified models:

GBR: learning rate = 0.01; maximum depth = 3, n estimators = 300.

XGB: subsample = 0.6, reg lambda: 50, reg alpha = 0.1, n estimators = 100, min child weight = 5, max_depth = 10, learning rate = 0.05, gamma = 0.1, colsample bytree = 8

XGBoost had a slightly better performance than GBR with an R^2 of 0.76 versus 0.75, an MAE of 42352.7 against 46769.5, and an RMSE of 82386.32 versus 83447.5.

C. Discussion

The ninth and last step of the KDD process is to document, or act, on the knowledge discovered during the data mining process.

The Irish housing market has been subject to high property inflation increases in the last 5 years. Much of this increase is attributed to many factors, such as housing supply needs and international migration. Moreover, in predicting housing supply needs for the Irish population the CSO sets a yearly requirement of 54000 new builds. [31] This was formerly estimated to total 30000 new houses per year from 2020 to 2030. [32] Clearly, macro-economic trends can be so unpredictable as such that in a period of 4 years the property supply needs nearly doubled.

Planning in such markets is often an exercise of compiling estimates that carry previous periods' risk. Evidently, the choice between constructing models with internal or external economic factors is primarily dependent on the size of the market and its agents.

This research attempts to study movements in the property market using a mix of internal and external pressures using machine learning. The choice of training features was motivated by the Irish economy being dependent on international fluctuations in supply and demand, specifically on a financial level. [33] Thus, financial market indices incorporate world and regional movements in all economic indicators and serve as comprehensive aggregations of economic influences.

The dominating cluster in our data has a mean price of 242720.53 and a std deviation of 88902.55. House prices within this range of the mean represents over 90% of our data. In this context, however, better results were obtained when combined with rather smaller observations of other clusters. This emphasises the macro viewpoint this research adopts.

The results of our research suggest a relationship between the Irish property price levels and movements in FTSE 100 and VWO Emerging Markets Index; the first one being a London Stock Exchange native, thus emphasising dependencies between the Irish and British economies. As for VWO, the link is theoretically unclear and requires further research. These international forces combined with unemployment, crime, net migration, new builds, and county offer a novel approach to modelling property price movements in Ireland.

V. Conclusions And Future Works

The study undertaken employed the use of multiple machine learning approaches, namely K-means clustering, support vector machines, stepwise regression, boosting, bagging, Random Forest, and stacked regression, to assess movements in house prices. The methods studied have shown promising results, particularly the tuned XGBoost and gradient boosting regressors.

Most of the approaches detailed in the literature focused on predicting house prices using granular data relating to the properties themselves, such as size, number of rooms, and neighbourhood.

However, the results of this study show that, using a novel approach and with a combination of multiple publicly available data, one can assess movements in housing prices with reasonable levels of accuracy.

In the face of increased costs of living and post-COVID economic recovery the results of this study can be used by stakeholders to take macroeconomic decisions which are of benefit to Ireland.

Future studies on this subject can take a variety of forms. Firstly, one can apply alternative methods of learning to the studied data. As observed in some studies [5] [7], artificial neural networks are a good candidate for this as the number of housing transactions already existent is sufficiently large, and it will only keep increasing over time.

Secondly, there may be other forms of macroeconomic data which could be relevant to housing prices, such as interest rates and consumer price indices. If the study were to be undertaken by state entities, it may also supplement this with economic indicators which may not be available to the general public.

REFERENCES

[1] Housing Europe, “The State of Housing in Europe 2023”.

[2] Central Statistics Office, “Residential Property Price Index January 2023,” [Online]. Available: https://www.cso.ie/en/releasesandpublications/ep/p-rppi/residentialpropertypriceindexjanuary2023/. [Accessed 16 October 2024].

[3] Central Statistics Office, “Residential Property Prices - Ireland and the EU at 50,” 17 October 2023. [Online]. Available: https://www.cso.ie/en/releasesandpublications/ep/p-ieu50/irelandandtheeuat50/economy/residentialpropertyprices/. [Accessed 16 October 2024].

[4] T. Alshammari, “Evaluating machine learning algorithms for predicting house prices in Saudi Arabia,” in *2023 International Conference on Smart Computing and Application (ICSCA)*, 2023.

[5] J. Kim, Y. Lee, M.-H. Lee and S.-Y. Hong, “A Comparative Study of Machine Learning and Spatial Interpolation Methods for Predicting House Prices,” *Sustainability,* vol. 14, p. 9056, July 2022.

[6] W. K. O. Ho, B.-S. Tang and S. W. Wong, “Predicting property prices with machine learning algorithms,” *Journal of Property Research,* vol. 38, p. 48–70, October 2020.

[7] T. D. Phan, “Housing Price Prediction using Machine Learning,” in *2018 International Conference on Machine Learning and Data Engineering,* Sydney, 2028.

[8] S. Lahmiri, S. Bekiros and C. Avdoulas, “A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization,” *Decision Analytics Journal,* vol. 6, p. 100166, March 2023.

[9] P. Rezazadeh Kalehbasti, L. Nikolenko and H. Rezaei, “Airbnb Price Prediction Using Machine Learning and Sentiment Analysis,” in *Lecture Notes in Computer Science,* Cham, Springer International Publishing, 2021, p. 173–184.

[10] A. Barlybayev, A. Sankibayev, R. Niyazova and G. Akimbekova, “Machine learning for real estate valuation: Astana, Kazakhstan case,” *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 35, p. 1110, August 2024.

[11] S. C. Yucebas, S. Yalpir, L. Genc and M. Dogan, “Price Prediction and Determination of the Affecting Variables of the Real Estate by Using X-Means Clustering and CART Decision Trees,” *JUCS - Journal of Universal Computer Science,* vol. 30, p. 531–560, April 2024.

[12] J. K. B. Byeonghwa Park, “Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data,” *Expert Systems with Applications,* vol. 42, no. 6, pp. 2928-2934, 2015.

[13] Q. Zhang, “Housing Price Prediction Based on Multiple Linear Regression,” *Scientific Programming,* vol. 2021, no. 1, pp. 1-9, 2021.

[14] K. H. Liow, “Dynamic relationship between stock and property markets,” *Applied Financial Economics,* vol. 16, p. 371–376, March 2006.

[15] J. Kakes * and J. W. Van Den End, “Do stock prices affect house prices? Evidence for the Netherlands,” *Applied Economics Letters,* vol. 11, p. 741–744, October 2004.

[16] M. Irandoust, “The causality between house prices and stock prices: evidence from seven European countries,” *International Journal of Housing Markets and Analysis,* vol. 14, p. 137–156, May 2020.

[17] V. Ceccato, “The impact of crime on apartment prices: evidence from stockholm, sweden,” *Geografiska Annaler: Series B, Human Geography,* March 2011.

[18] R. Sanchis-Guarner, “Decomposing the impact of immigration on house prices,” *Regional Science and Urban Economics,* vol. 100, p. 103893, May 2023.

[19] L. Gan, P. Wang and Q. Zhang, “Market thickness and the impact of unemployment on housing market outcomes,” *Journal of Monetary Economics,* vol. 98, p. 27–49, October 2018.

[20] G. P.-S. P. S. Usama Fayyad, “From Data Mining to Knowledge Discovery in Databases,” *AI Magazine,* vol. 17, no. 3, pp. 37-54, March 1996.

[21] Central Statistics Office (Ireland), “HPM09 - Residential Property Price Index,” [Online]. Available: https://data.cso.ie/table/HPM09. [Accessed 25 October 2024].

[22] Central Statistics Office (Ireland), “CJA01 - Recorded crime incidents,” [Online]. Available: https://data.cso.ie/table/CJA01. [Accessed 25 October 2024].

[23] Central Statistics Office (Ireland), “PEA15 - Annual Population Change,” [Online]. Available: https://data.cso.ie/table/PEA15. [Accessed 25 October 2024].

[24] Central Statistics Office (Ireland), “F1009 - Annual Estimated Net Migration per 1000 of Average Population,” [Online]. Available: https://data.cso.ie/table/F1009. [Accessed 25 October 2024].

[25] Central Statistics Office (Ireland), “MUM01 - Seasonally Adjusted Monthly Unemployment,” [Online]. Available: https://data.cso.ie/table/MUM01. [Accessed 26 October 2024].

[26] Central Statistics Office (Ireland), “FY056B - Rates for Labour Force Participation and Unemployment,” [Online]. Available: https://data.cso.ie/table/FY056B. [Accessed 26 October 2024].

[27] Central Statistics Office (Ireland), “NDQ05 - New Dwelling Completion,” [Online]. Available:

https://data.cso.ie/table/NDQ05. [Accessed 26 October 2024].

[28] Yahoo Finance, “Markets: World Indexes, Futures, Bonds, Currencies, Stocks & ETFs,” [Online]. Available: https://finance.yahoo.com/markets/ . [Accessed 25 October 2024].

[29] G. James, D. Witten, T. Hastie, R. Tibshirani and J. Taylor, An Introduction to Statistical Learning: with Applications in Python, Springer Nature, 2023.

[30] A. Ahrens, E. Ersoy, V. Iakovlev, H. Li and M. E. Schaffer, “An Introduction to Stacking Regression for Economists,” *Studies in Systems, Decision and Control,* vol. 429, pp. 7-29, 2022.

[31] Central Bank of Ireland, *Economic policy issues in the Irish housing market - Quarterly Bulletin 3,* Dublin: Central Bank of Ireland, 2024.

[32] A. Bergin and A. García-Rodríguez, “Regional demographics and structural housing demand at a county level,” The Economic and Social Research Institute, Dublin, 2020.

[33] Egan, M. K. Paul and C. O'Toole, “How supply and demand affect national house prices: The case of Ireland,” *Journal of Housing Economics,* vol. 65, 2024.