National College of Ireland

Stacking Ensemble Approach towards Predicting Irish-Real Estate

MSc Research Project
Msc_DAD_A_JAN24O

Diepesh Manish Patel
Student ID: x23197609

School of Computing
National College of Ireland

Supervisor: Prof . Aaloka Anant

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Dipesh Manish Patel
Student ID:	x23197609
Programme:	MSC_DAD_A_Jan24O
Year:	2024
Module:	MSc Research Project
Supervisor:	Prof Aaloka Anant
Submission Due Date:	12/12/2024
Project Title:	Stacking Ensemble Approach towards Predicting Irish-Real Estate
Word Count:	5661
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Dipesh Manish Patel
Date:	28th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Stacking Ensemble Approach towards Predicting Irish-Real Estate

Dipesh Manish Patel
X23197609

Abstract

Through visual analytics, Study aims to explore whether signs of property market bubble formation have become evident, and to create a pragmatic predictive model of Irish real-estate valuation. Acknowledging the challenge of the market, this study is a holistic approach aware of the many limitations of traditional investigation that creates a one-sided distinction between the factors influencing property values.

This study utilizes a sequencing ensemble technique that encompasses Random Forest, LightGBM and XGBoost for improved prediction accuracy at the level of machine learning algorithms. With an R^2 score of 0.99, this model proves to be a sturdy predictor of valued property; This study utilizes elements such as demographic data, economic indicators, and real estate transaction records to provide an extensive overview of the market. The data also show a 70% increase in property prices, which in turn highlights a widening gap of property prices against average salaries, leading into market corrections. These findings will learn important implications on researchers.

Keywords: Real Estate, Price Prediction, Ensemble Method, Random Forest, XGBOOST, LightGBM

1 Introduction

In rapidly evolving markets, such as Ireland, the accurate prediction of real estate prices has emerged as a crucial challenge for homeowners, investors and policymakers. One of the most distinctive facets of the Irish real estate landscape is its propensity for cyclical evolution, influenced by both national economic cycles and international FDI and global paradigms Wandhe et al. (2023). Its complexity makes it an interesting study in future market trends and robust forecasting models. The present research focuses towards identifying a real estate bubble tendency and harnessing an effective predictive modeling mechanism to Irish property valuation through visual analytics. Although there have been significant advances in machine learning and data analytics, limited exploration of these approaches used holistically are applied to the Irish real estate market. There have been only a handful of attempts to cross various dataset types, including demographic data, economic indicators, and records of real estate transactions, to form a more multidimensional analysis. Moreover,according to Singh et al. (2023) the available evidence on the effectiveness of various data balancing techniques and outlier removal to enhance predictive accuracy is limited. At the same time, they left gaps in the model, stretching it to work outside of its intended use case: a second key dataset of this type presents itself

to allow for a more comprehensive and accurate approach that is suited to the nuances of the Irish market.

Therefore, this research seeks to provide valuable insight on the Irish property market by responding to these objectives through the design and development of a predictive model, enhanced by analyzing the key factors that influence the value attributed to a property. These findings will not only assist stakeholders making decisions but also address significant gaps in the existing literature, laying the groundwork for future research pathways in the realm of real estate market dynamics.

Research question

How effective staking ensemble predictive model approach for forecasting the Irish real estate price. Sub-research question:- using visual analytics try to detect real-estate market bubble in residential property.

To address the above question the prediction of real-estate prices was widely explored but as mentioned in the report even minor changes in data or feature can make a significant difference in forecasting outcome. Ireland is famous in Europe and outside Europe about its tax heaven status which attracted the investors the publicly available dataset for government web portal is in finite range in terms of features and their missing value and with the combination of other data set and removing prominent feature makes this report unique.

Research Objective

1. Determine the correlation between datasets for better understanding.
2. Visualizing the data with the correlation of the features for better understanding of the Irish real estate market with nations economic growth
3. Evaluate if there is any presence of a real estate bubble by market by comparing the visualization of data.
4. Use machine learning model to predict the real-estate price for best accuracy.

2 Related Work

Providing an overview of the existing literature on real estate price prediction, in terms of techniques used and results obtained, this section presents the current report's methodology for predicting Irish real estate prices through advanced machine learning and data analytics techniques, and how it compares to the other relevant using the works identified in this overview.

2.1 Employment and Culture Influence on Real Estate Prices

According to Agnew and Lyons (2018) economic analysis of the effects on Irish residential property values of employment growth triggered by foreign direct investment (FDI) Theirs found that residential property prices for every 1,000 jobs that are created by FDI, on average residential property prices after a 1–2 year lag increased by about 2%. Similarly, Moro et al. (2011) estimated a 0.4% to 0.6% increase in the price of a real estate located near cultural properties and historical monuments based on the answers obtained from the residents. The implications of these results is that employment data and societal variables are relevant when creating predictive models for Irish property markets. .

2.2 Economic Cycles and Affordability and Irish Real Estate

According to Corrigan et al. (2019a) Housing finance in Ireland (2019) identifies that over 30 percent of tenants in Dublin currently pay more than 30 percent of their income on (gross) housing costs, and makes the case for necessary policy reforms. Analyzing Airbnb rental data on short-term leases in Dublin,Lima (2019) tackled what is still a constant problem in this area, which relates to the specifications of datasets, because it was not possible to identify who the actual tenant was (which indicates that not everything in that universe of rentals was aiming to settle as a short-term tenant). These results point to the importance of incorporating inflation rates and minimum wage information to address affordability questions in models' predictions.

2.3 Price Prediction Models

Corrigan et al. (2019a)In New Zealand ANN was compared with hedonic pricing models with (2004). ANN gives better performance than multiple linear regression with R2 value¡75% in case of non-linear relationship. Similarly, Agnew and Lyons (2018) and used Random Forest models to predict property prices in Arlington County, Virginia, leveraging properties size and distance from city centers as features superior than other models.

Rana et al. (2020)A dataset was used with over 113,000 records compared a number of models using Random Forest, XGBoost and Support Vector Regression (SVR),(2020) It gave Highest Classification Accuracy as Random Forest (94%) and XGBoost (63%) gave lowest accuracy. Madhuri et al. (2019a) used boosting methods to predict real estate prices, is better than traditional linear regression (85%), without feature engineering.

Singh et al. (2023) used exploratory data analysis with machine learning algorithms (gradient boosting and Random Forest) and achieved 87% success rate in predicting the price of the vehicle. But their study didn't take into account market fluctuations over time. For example, Alshammari (2023a)found that Random Forest was (89%) outperformed by decision tree regression and linear regression when predicting house prices in Saudi Arabia. Corrigan et al. (2019b) that 91% accuracy was achieved from model stacking, but at a computational cost. Wandhe et al. (2023) 93%The study successfully, but is something related to sparsity in the data and no comparison with other advanced methods.

Peng et al. (2019) provided a prediction in the context of second-hand housing price prediction, with 92% accuracy using XGBoost, however, did not mention feature importance or interpretability.

2.4 Literature Niche

The missing piece in the area of real estate predictive modeling is the incorporation of numerical and temporal data, improved interpretability, and scalability for real-time applications. Gotten attention on hybridization in existing work tends to be on a more low-level and proposed frameworks remain largely un-choked; more advanced ensemble systems like stacked hybrid models, and methods to reduce imbalance and sparsity in data are largely unaddressed. Nonetheless, most of these approaches either do not combine all the components or use trained weights which makes the interpretation of dynamics and mechanisms difficult, and the operation even for a wide range of markets requires further exploration, and there is indeed a need for more work towards fully integrated,

interpretable and scalable model aimed at real-world markets that are dynamic themselves.Wandhe et al. (2023),Wandhe et al. (2023),Alshammari (2023a)

2.5 Novelty

In this study propose a more holistic framework to predict the price of Irish real estate by enhancing machine learning models with diverse datasets comprising demographic data, economic indicators, and records of real estate transactions. It uses a stacking ensemble of Random Forest, LightGBM and XGBoost to improve accuracy compared to standard studies. They even make use of visual analytics, feature engineering and data balancing techniques to tackle problems such as data sparsity, complex interactions, and model overfitting. Incorporating temporal dynamics and cultural influences, it offers a multidimensional perspective on the elements influencing Irish real estate prices, raising the bar for predictive modeling in this arena.

Table 1: Summary of Literature Review on Real Estate Price Prediction

Author(s)	**Focus of Study**	**Key Findings**	**Research Gap**
Agnew and Lyons (2018)	Impact of FDI on Irish real estate	FDI-driven employment growth increased property prices by 2%, with a lag of 1–2 years	Lack of integration with other economic factors
Moro et al. (2011)	Cultural heritage's influence on real estate	Historical landmarks increased nearby property prices by 0.4–0.6%	Limited exploration of cultural impact on broader markets
Corrigan et al. (2019a)	Housing affordability in Ireland	30% of Dublin tenants spend over 30% of income on housing costs	No dynamic or regional data analysis
Lima (2019)	Short-term rental affordability using Airbnb data	Analyzed affordability using short-term rental data for Dublin	Limited dataset scope; no inclusion of long-term trends
Limsombunchai et al. (2004)	Comparative study of ANN and hedonic pricing	ANN outperformed hedonic models with R^2 above 75% for non-linear data	Focused only on small datasets and limited features
Soibelman and González (2002)	Random Forest for U.S. property prices	Random Forest captured non-linear relationships better than other models	Limited geographic scope and feature diversity
Rana et al. (2020)	Comparison of ML models for price prediction	Random Forest achieved 94% accuracy; XGBoost performed moderately	Middle overfitting and lack of interpretability
Madhuri et al. (2019a)	Regression methods for house prices	Boosting methods improved accuracy (85%) over linear regression	Limited feature engineering and scalability
Singh et al. (2023)	Machine learning for automobile pricing	Gradient boosting and Random Forest achieved 87% accuracy	Lacked dynamic market and temporal data analysis
Alshammari (2023a)	ML models for Saudi housing prices	Random Forest achieved the highest accuracy (89%)	No inclusion of macroeconomic variables
Alshammari (2023b)	Stacking and ML techniques for price prediction	Stacking achieved 91% accuracy; SVM and linear regression performed lower	High computational cost; runtime inefficiency
Wandhe et al. (2023) (2023)	Hybrid CNN and NLP for price prediction	Achieved 93% accuracy using text and numeric data	Data sparsity issues and lack of comparison with advanced methods
Peng et al. (2019)	XGBoost for second-hand housing prices	Ensemble methods achieved 92% accuracy	Lacked feature importance analysis and interpretability

3 Methodology

In this study Knowledge Discovery in Databases (KDD) methodology. KDD is an ideal choice for this task because it provides a comprehensive framework for extracting useful knowledge from large volumes of data. The KDD process is particularly well suited for handling complex datasets and uncovering hidden patterns, relationships, and trends that can be used to make informed decisions and generate valuable insightsSoibelman and González (2002) Figure 1.

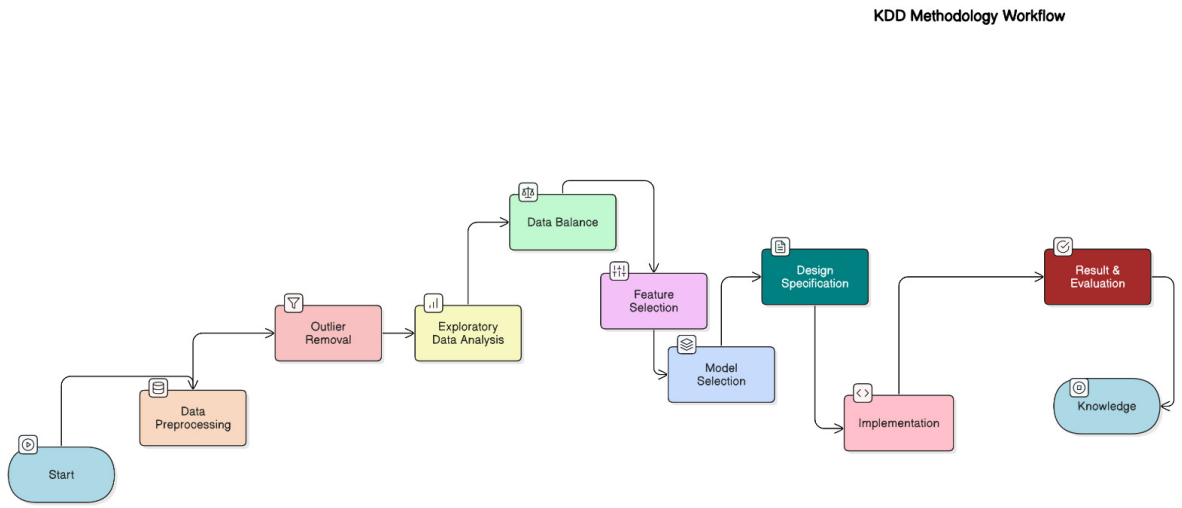

Figure 1: KDD Methodology

3.1 Start

In this section study ensured that all data loaded successfully and ready Majority of the property sales data were from same source with one difference each data set contain the property sales records of only that particular year and all the other features were same and all of the data were encoded and cant be loaded using normal pandas command.to save time study created a list containing all the similar data set from 2017 to 2023 then study used for loop and called the list and used encoding cp1252 to transfer that records into data frame because it was a large amount of data as a precaution Study put low memory as false to avoid issue with mix data type.as an example price feature had special character and numerical value.

Before second merged study converted the data of sale feature from string and split the feature into 2 features years and month to merged it with salary data set after that .to merged the Salary data set first I filtered specific features and records and removed unwanted records and feature from Salary data set so I can merged them based on common feature county and year as a result after second merged new feature is added to data frame as salary. [1] [2] [3] [4]

[1] Workplace Relations - Ireland
[2] Property Price Register - Ireland
[3] Inflation Statistics - Europe
[4] Central Statistics Office - Databases

3.2 Data Preprocessing

After first Merged validation the data types of the records and null values where checked based on that few featured where removed for having more than 60 percent null.

In the second merge validation I filtered the features and records to satisfy the condition of merge and after that I validate using sample rows where random 5 rows were shown.

In third merged validation checked the data type again to be sure that the changes that were made in first 2 merge were there like splitting of date feature into month and year. Also checked for any duplication and removed those records.

In terms of reliability test I checked the duplicate value and removed the duplicate records. .

3.2.1 Conversion of Categorical Feature into Numerical

So in this section we performed one hot encoding because converting categorical features into boolean helps machine learning model to evolved further and will get better accuracy Alshammari (2023a).So in this study performed one hot encoding on "County "feature.

3.2.2 Merging Datasets

So in this section , all the features were merged into single datasets for further processing.

3.3 Outliers Removal

To check if the real estate price has too much price fluctuation study uesd Z-score methods and using the Box plot to identify outliers is an effective technique according to Corrigan et al. (2019b). so in next sub-section disussed a outliers which present in Price features.So in this study used Price as a predictor.

3.3.1 Before Removing Outliers

After applying the Z-score as well as the Box plot method following figure demonstrate the outliers that present in Price feature.the 2.0 is 2 million . Figure 2.

3.3.2 After Removing Outliers

In scenarios such as predicting real estate pricing accuracy, outliers can hugely skew statistical analyses and models' real performance because extreme values could reflect unusual market circumstances but not overall trends. To solve this problem, the Outlier analysis used Interquartile Range(IQR) method which is a proper statistic technique used to find and eliminate outliers as it doesn't get impacted by the extremities in the data.Lima (2019) This method helps ensure that the dataset will correspond to the actual central distribution of the data, thereby improving model performance and contributing to the best analysis, it is more reliable and interpretable. The dataset was filtered using the IQR method by removing outliers (extreme points), finally the obtained dataset was cleaner and more concise.Figure 3.

Below formula used for removing outliers in data.

$$\text{IQR} = Q_3 - Q_1 \tag{1}$$

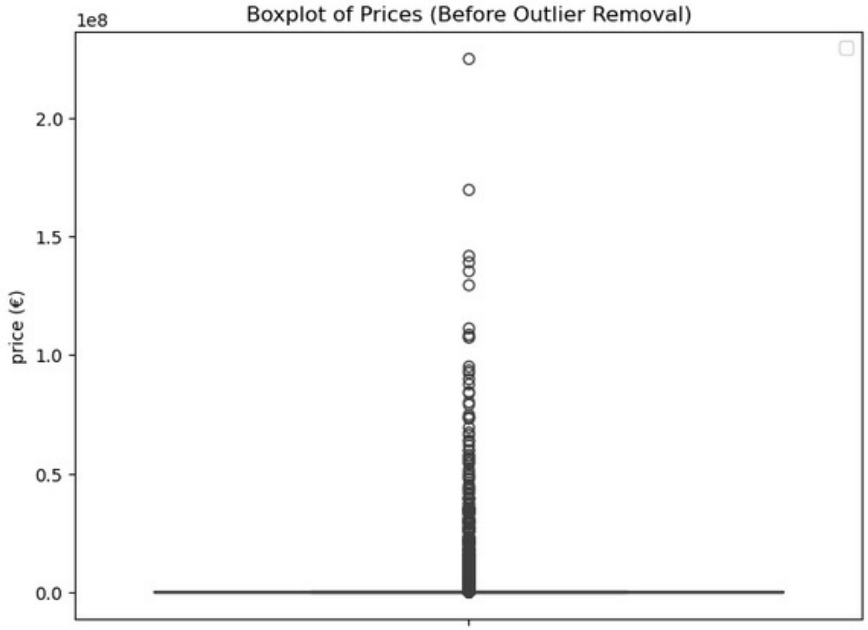

Figure 2: Before Removing Outliers

Based on the visualization of boxplot the value of Q1 where it represent the 25 percent of the real-estate price which is approximately 1,60,000 euro and Q3 which represent 75 percent of real-estate price which is approximately 3,60,000 euro which makes the IQR property price value approximately 2,00,000 euro.

3.4 Exploratory Data Analysis

In this section study trying to check the correlation and between the features and the possibility of real-estate bubble based on county. After satisfying with data outcome, the study performed Exploratory Data Analysis (EDA) on the data after removing outliers to get a clear result. As shown in Figure 4. the price started from 220,000 and ended at 375,000 with a small fluctuation in sale. During the term from 2017 to 2024 the prices were increased by 70.40[5]

[5]Interquartile Range - Wikipedia

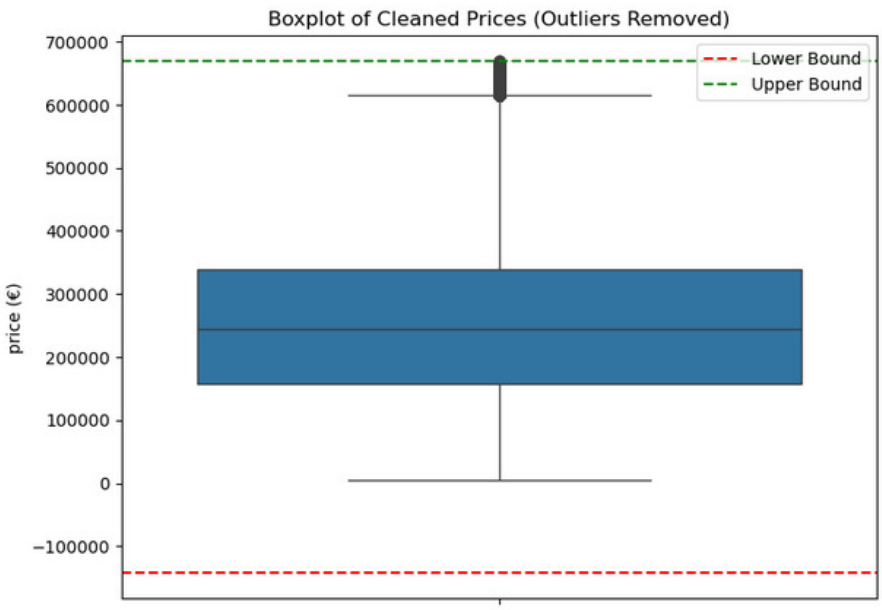

Figure 3: After Removing Outliers

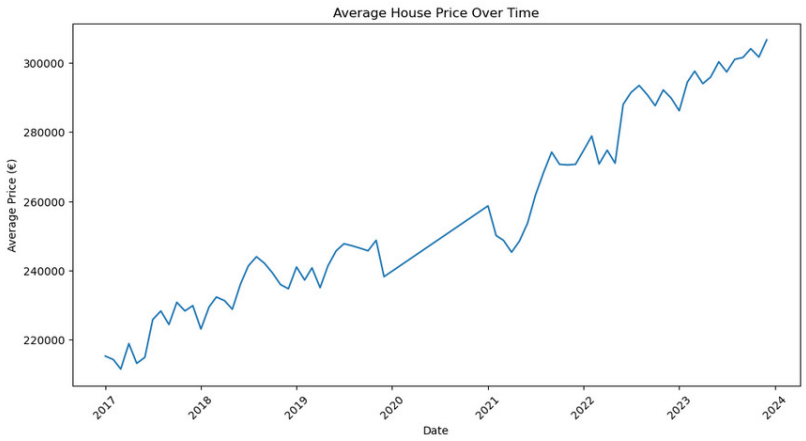

Figure 4: Line chart of Average Housing Price

To get a more detailed analysis Study made a pie chartFigure 5 of 4 Irish regions where Region Leinster which contains 56.8% off total houses sold as it has an extra number of counties than others, is the 2nd most houses sold in Ireland from 2017 to 2023 where Region Muster with 26.5% half of the Leinster region. As shown in Figure 5 region Connacht and ulster were last with 11.6 and 5.2 house sails to get to know if there is any risk of the real estate bubble. Study filtered the region Leinster data to find the top 5 counties with most house sold. The rank was as follows County Dublin, County Kildare, County Meath, County Wexford, County Wicklow.

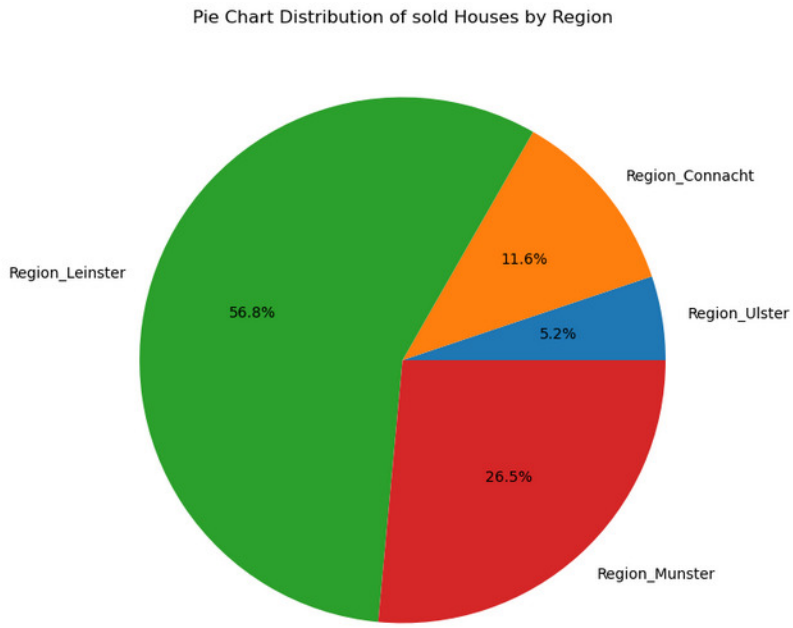

Figure 5: Pie chart Distribution of sold Houses by Region

After performing EDA study found out that the data was highly unbalanced Figure 6because of that data set might create high bias which can also generate poor accuracy to balanced the data study used underdamping method to balanced the data on Vat exclusive feature as shown in figure 7 Vat exclusive is a Boolean feature with 2 option yes=1 or no=0 .to perform underfitting study used random method to select the same number of variable with value 1 from Vat exclusive.

In terms of business it gives the clarity of property price and a basic property info if the property is old or new. The VAT registered business often do business in VAT property as then can typically recovered the VAT which helps business accurately calculate profit without any distribution of VAT

Below formula used for Calculating VAT.

$$\text{VAT exclusive price} = propertyprice * (1 + VATrate) \tag{2}$$

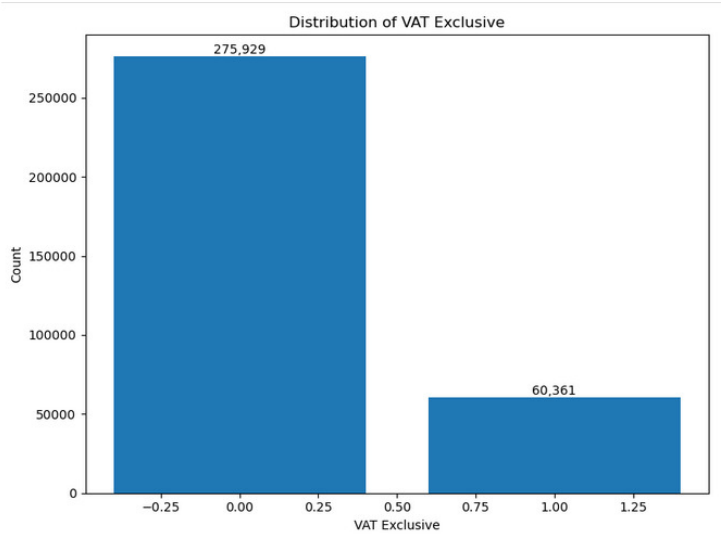

Figure 6: Distribution of VAT in 0 and 1

3.5 Data Balance

As per the study to avoid the biased output study used VAT exclusive feature to balanced the data as in data set the resiling property data is more than newly selling property to get the better accuracy I used under sampling and balanced the data were the I randomly selected the same number of records of second hand property as newly developed property.

The balanced dataset distribution for the VAT Exclusive featureFigure 7. In both categories, 60,361 observations were used after balancing the twin sample. Having this distribution of the classes will make the model fair because there will be no biased cannot be caused because of how the classes are not balanced out so that during the training phase the model should not be weighted to any one of them either. This is followed up by the balancing methods being applied, such as oversampling, undersampling, or the generation of synthetic data (SMOTE), so that every class present in the dataset is represented correctly. Thus, extracting features from the training dataset dynamically removes the overfitting or underfitting, and thus generates the predictive model with better generalization, making the predictions more accurate. This part of the model is strong and unbiased is very important when we working on heavily unbalance data.

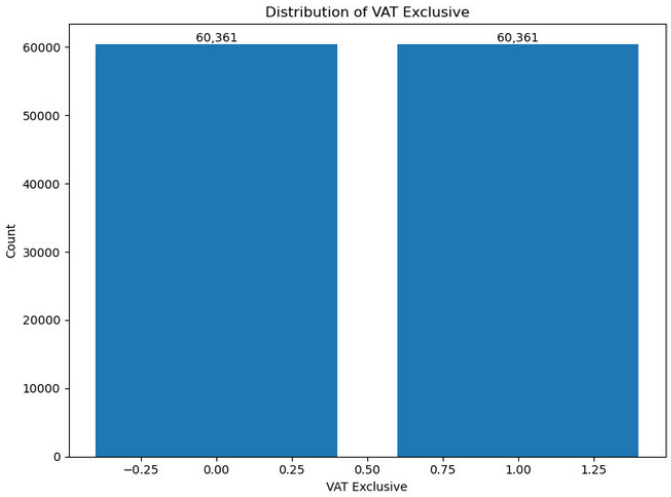

Figure 7: Balanced Distribution of VAT in 0 and 1

3.6 Feature Selection

The correlation heatmap draws crucial insight on the relationships between the features says by researcher Madhuri et al. (2019a) , as well as their correlation to the target variable, price (€), to inform feature selection for this study. Direct correlations, including Salary (€) at 0.49 and Region Leinster at 0.36, position them as some of the most significant predictors of property value. The importance of these factors in explaining price variations is also witnessed in seasonality (represented by Year correlation of 0.20) and economic indicators such as Inflation correlation of 0.18. With features like VAT Exclusive (0.18 correlation) it emphasizes taxes on property pricing.

In contrast, the Not Full Market Price (-0.16 correlation) and Description of Property (-0.18 correlation) features are relatively weakly correlated with the target variable, highlighting their minimal contribution to prediction. Furthermore, Multicollinearity is being treated by looking at pairs of closely related features, like those Year with Unique ID (0.98) and Inflation with Min Wage (0.85) This is a technique to reduce redundancy where they keep only a single most informative feature from these pairs. The dataset is further refined through feature selection based on strong or moderate correlations, while taking care of multicollinearity, leading to a more focused set of predictors that enhance model explanatory power and interpretability. Figure 8

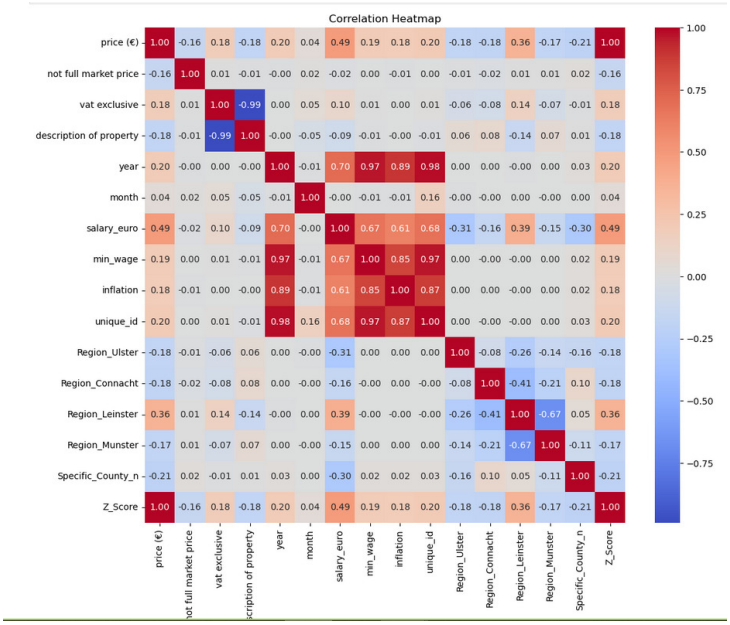

Figure 8: Corelation Matrix

3.7 Model Selection

Due to the size of the dataset and its complexity, the study selected a stacking ensemble approach as the predictive model. Bagging is an advanced machine learning technique that generates multiple base models and combines them in some way to improve accuracy and robustness. The ensemble in this study uses Random Forest, XGBoost, and LightGBM as base models, and a regressor meta-model that learns from the predictions set generated by these base models. By harnessing the strengths of each model, i.e. Random Forest for handling nonlinear relationships, XGBoost for optimizing using gradient boosting and LightGBM efficiency for larger datasets. Past studies strongly back the choice to stacking. For example,Alshammari (2023a)reached 89% accuracy in predicting housing prices using Random Forest models, while Peng et al. (2019) established XGBoost was very effective when forecasting second-hand housing price with an accuracy as high as 92%. Additionally,Madhuri et al. (2019a)proved the stability of Random Forest and XGBoost with a similar prediction accuracy of 80% for automobile prices. All of these studies report accuracies of greater than 85% and make these models a good choice for price prediction tasks. The stacking approach maximizes the performance by balancing out the individual model shortcomings with the strengths of other models. As depicted in Figure 9, the stacking ensemble is specifically designed to optimize predictions in a systematic manner. Both empirical evidence and the needs of the dataset further support this choice, leading to strong and valid predictions for the study.

4 Design Specification

Thus, the design specification for this study presents a stacking ensemble architecture that is a sophisticated machine learning approach which integrates predictions from several base models to improve accuracy, robustness, and generalizability for the price prediction. It all starts with a labeled dataset with input features and target variables. The dataset

is preprocessed by treating Null values, removing outliers, and balancing the data so that all the classes are well represented. Is crucial to massively reduce model bias and enhance prediction reliability.This is a simple stacking ensemble with three base models: Random Forest (RF), LightGBM, and XGBoost. All models are chosen as per their individual strengths and established predictive performance on similar prediction tasks. Random Forest can handle non-linearity and complex interactions among features, making it useful for complex data. Such an intelligent estimator is largely based on the efficiency of its computation, allowing LightGBM to work with very large data and imbalanced data without sacrificing accuracy. XGBoost is a gradient boosting framework that has been widely used in various data science competitions due to its ability to achieve high accuracy. The dataset is processed independently by each base model, predicting output using its strength in algorithm. The predictions of these base models are routed to an ensemble model which is the meta-learner. This meta-model combines all the predictions made by the base models and then learns how to combine them properly before producing the output. The ensemble can learn from the strengths of each base model so that their combined prediction is more accurate and robust than any single model The stacking ensemble addresses weaknesses of individual models and improves overall predictive performance through the entire process, making it a powerful method for capturing complex relationships in data says by Corrigan et al. (2019a). This architecture is well motivated by the success of these models in past experiments. Random Forest, for example, reached an accuracy of 89% in predicting housing prices in Saudi Arabia Alshammari (2023a) and XGBoost reached an accuracy of 92% in predicting second-hand housing prices in Chengdu (Peng et al., 2019). Additionally, Moro et al. (2011) established that Random Forest and XGBoost are robust predictive models of automobile prices with an accuracy exceeding 80%. The consistent performance of these models in price prediction tasks confirms the reliability and effectiveness of these models as an ideal candidate to be used in this study. The ensemble is further complemented with LightBGMs computational efficiency, enabling optimal handling of large datasets with very little overhead.There are certain advantages to stacking ensemble architecture. By combining models with different strengths, it enhances predictive accuracy, leading to improved performance on complex datasets. It increases the robustness of models, leadding them to be less prone to overfitting and more useful on unstimulated data. Furthermore, the average of different baseline models allows for interpretability, as we can analyze the impact of each training through the contribution of each model towards the baseline contribution of the final prediction. By following this iterative process, the model is more likely to be appropriately designed to tackle the intricacies of the dataset while providing robust and accurate predictions.The design diagram in Figure 9 demonstrates a straightforward construction for our architecture where the labeled dataset flows through the base models to provide an output, which drives our ensemble model to the final prediction. Such design guarantees that the stacking ensemble architecture is not only overall robust and accurate, but also well suited to tractability of the ever-changing environment of price prediction tasks, making it the best solution for the needs of this study Madhuri et al. (2019b).Figure 9

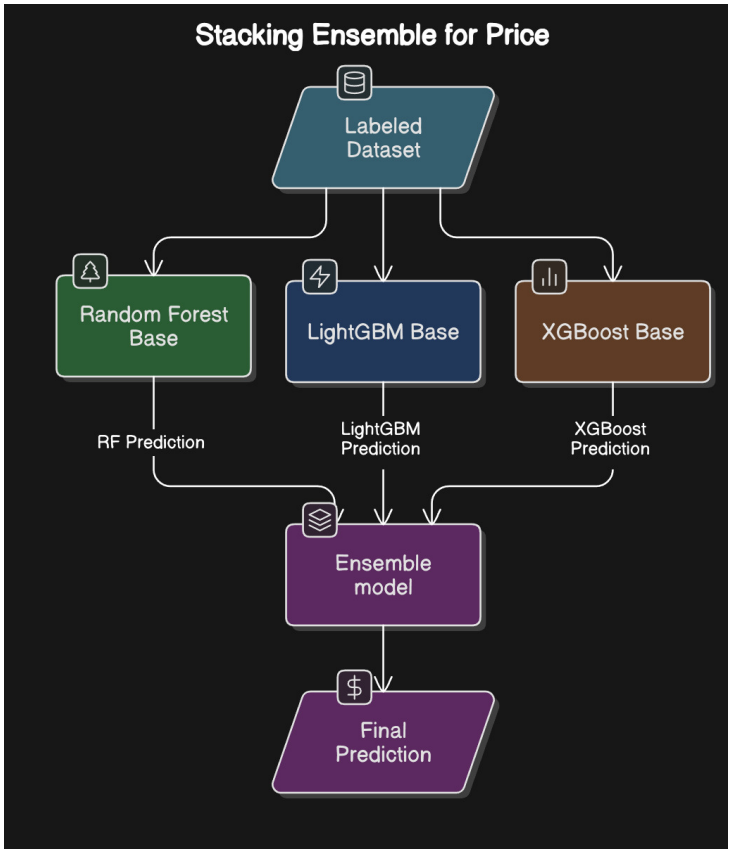

Figure 9: Stocking Ensemble for Price

5 Implementation

Combining the models together was done with the stacking ensemble model The price prediction model used Python, advanced libraries for preprocessing [NumPy, Pandas], visualization [Matplotlib] and machine learning libraries [RandomForest, XGBoost, LightGBM]. To de-feature the unbalanced classes SMOTE was used to perform a create a balanced dataset. The data was divided into80% training and 20% testing, sufficient data for model training and testing. The target variable which is price was identified, and the features were Preprocessed.

In this case, the base models — Random Forest, XGBoost and LightGBM — were trained individually, and they were initialized with the same hyperparameters. We kept n estimators = 100 for each model as a good trade-off between performance and function with the least computational overhead, and Random State = 16 to ensure reproducibility. Predictions generated from the base models were then used as input features to the ensemble model implemented by Random Forest with n estimators = 50 and Random State = 16 for aggregation of predictions. Final predictions were obtained through an ensemble model, and performance was evaluated using MSE and R^2 score to ensure robustness across performance metrics.Figure 11

To evaluate the potential existence of a real estate bubble, the study concentrated on the two counties with the most number of house sales. Looking at year-on-year growth rates for the property market revealed a mind-numbing graph Figure 10 that showed Dublin properties growing significantly from 2017 to 2018 while salaries (and rents) stayed

around 30-35%. On the contrary, Kildare reflected a balanced dynamic, with property prices and salary less divergent, therefore being more affordable. The minor downtrend in property prices for both counties, probably from filtering out outliers, suggested that the trends were caused by supply–demand imbalances, not that of a real estate bubble.

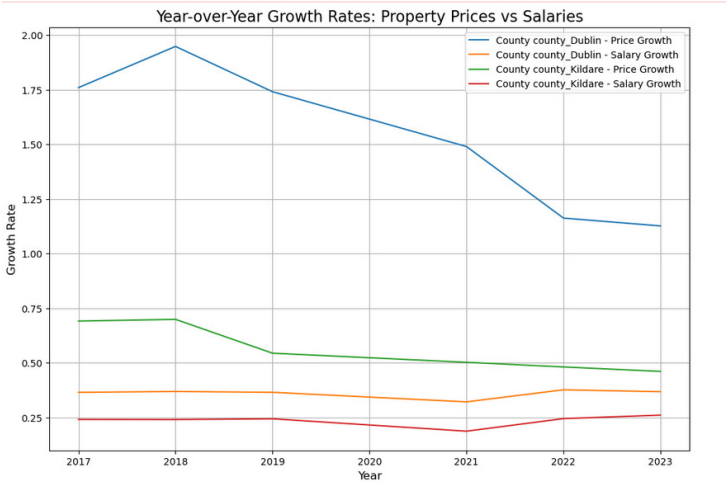

Figure 10: Property price vs Salary per county

In order to prove the significance of data balancing even more, the dataset was re-analyzed in the study without balancing. This step underscored the vital importance of balanced datasets in improving model robustness and prediction accuracy. This is even more important with very complex datasets where overfitting is likely to occur.Figure 11

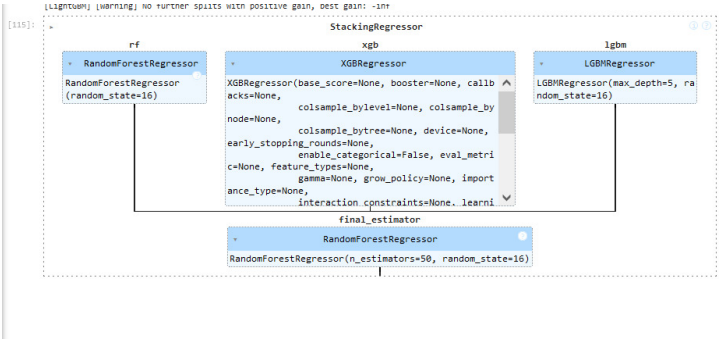

Figure 11: Implementation of Ensemble Model

6 Result and Evaluation

After performing ensembled model with the data balancing and removal of outliers through under fitting the R2 score with 0.99999 , almost perfect fit to test data which can translates to model having equivalent to 99% accuracy as with the Mean Square Error(MSE) of 941.89 which indicates the difference between the predicted value and actual values. This result showed that how effective ensembled method was on given data.Study also evaluated the base model for further research where base model Random Forest was having higher accuracy of R2 score of 0.99 as training and testing data set but

the MSE score difference between training and testing data with MSE 56.65 and testing MSE 427.75 but the gap was minimal not to be considered as overfitting.the XGboost and LightGBM are having same R2 value 0.99 but different MSE value with minimal differences hard to consider as overfitting.

The following figure Figure 12 shows how predicted values and real values of the target, the price, in the test phase look like in the same plot. The black line is the diagonal line (y = x). Such a high correlation implies that the expected values from the model are very close to the original values. These results emphasize that the stacking ensemble model is a significant improvement as it consolidates Random Forest, XGBoost, and Light GBM components, drawing upon the strengths of every base structure to produce accurate prediction. The lack of major outliers or deviations indicates the model is robust, and that the preprocessing steps, such as removing outliers and feature selection had a vital influence in enhancing performance. Furthermore, the model's ability to maintain accurate predictions across different price ranges further confirms the suitability of the adopted stacking ensemble modeling technique and highlights its generalizability to unseen data. In this plot as an example we see the model works well predicting prices in general, which shows that we can trust the model as a reliable and accurate tool. Figure 12

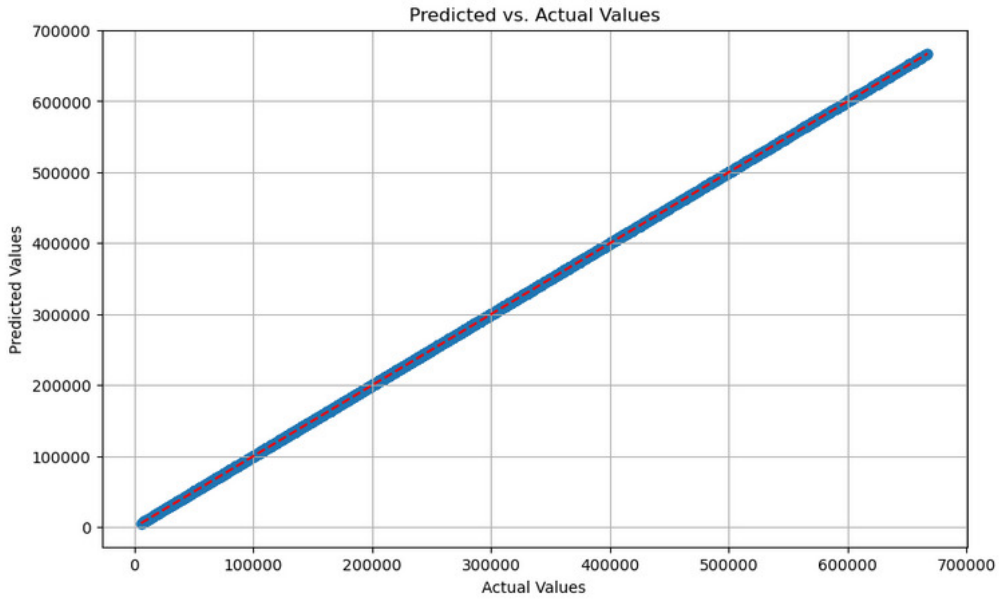

Figure 12: Actual vs Predicted

6.1 Expriment 1 : Using Random Forest

This model performed very well with the MSE values at a great value for train set (56.66) and test set (427.76). Both training and testing R^2 scores were almost perfect (0.99999); the model captured most of the variability of the data. For that reason, the performances of Random Forest were significantly good with very little overfitting or underfitting Table 2.

Table 2: Random Forest Model Performance

Metric	Result
Training Set MSE	56.65558833459786
Training Set R^2	0.9999999966819625
Test Set MSE	427.7558860278028
Test Set R^2	0.9999999744975391

6.2 Experiment 2 : Using XGBOOST

XGBoost was still not bad, but the MSE valuse were higher than Random Forest: 1,250,289.06 for the training set, 1,292,988.76 for the test set. R^2 scores were also very high, on the train/test was 0.99992. This means XGBoost worked well, but it can be improved further through hyperparameter tuning. Table 3

Table 3: XGBoost Model Performance

Metric	Result
Training Set MSE	1250289.0594796592
Training Set R^2	0.9999267744130875
Test Set MSE	1292988.7579121475
Test Set R^2	0.9999237484841505

6.3 Experiment 3 / Using Lightgbm

Similar to Random Forest and XGBoost, the LightBGMs achieved awesome performances, with a training MSE of 119,358.79, and a test MSE of 1,042,219.54. drop in R^2 scores, attaining 0.99995 for Train and 0.99994 for Test, slightly lower than Random Forest. These results indicate that LightGBM cope well with the dataset given its large size while still being able to deliver relatively high accuracy on predicting target values. Table 4

Table 4: LightGBM Model Performance

Metric	Result
Training Set MSE	119358.78887785838
Training Set R^2	0.999946741405122
Test Set MSE	1042219.539499897
Test Set R^2	0.999939884083736

6.4 Final Result

So well, in fact, even the Ensemble Model, stitching the predictions from the base models (Random Forest, XGBoost, and LightGBM) in a stacking architecture, performed with an amazing accuracy. Test dataset has a low MSE of 941.89, indicating that prediction error is really low. Additionally, the R^2 score was substantially close to 1, as shown in our example, indicating that the model captured most of the variability of the said data (R^2 = 0.9999999449). Table 5 By leveraging the complementary strengths of the individual

base models, the ensemble model outperforms any of the individual base models. The base-models alone gave good predictions, however the meta-model from the stacking ensemble architecture was used to aggregate these outputs for a better final prediction. This result verifies that the ensemble technique successfully minimizes the prediction error and provides high accuracy, which makes it a great fit for complicated tasks results, such as prediction of real estate price prediction.As mentioned previously, by creating an ensemble model one is able to take advantage of combined predictions (Tables 2, 3, and 4). This choice of architecture was not only due to the advantages of having multiple markers, but also due to the generalizability and robustness of the architecture, making it a great candidate for this study. This performance indicates the strength of stacking as a state-of-the-art machine learning method for price forecasting tasks.

Table 5: Ensemble Model Performance

Metric	**Result**
Test Set MSE	941.8906553070012
Test Set R^2	0.999999944906191

6.5 Discussion

Table 5 lists the performance of the stacking ensemble on real estate price prediction showcasing the power of using a stacking ensemble. So, the Model's MSE is only 941.89 which shows very less error in the prediction done on model andan R^2 score of 0.9999999449 indicates that this model explained almost all the variance in the target variable. These results demonstrate the enormous improvement using the stacking ensemble architecture; utilizing the diverse strengths of the base models(Random Forest, XGBoost and LightGBM).

We have seen that the ensemble model achieved higher accuracy than all the base models and also gave higher error reduction. For example, Random Forest gave a high accuracy on its test set, yet had an MSE of 427.75. Similarly, XGBoost and LightGBM, whose test set MSEs were 1,292,988.76 and 1,042,219.54, respectively, also indicated that individual models did not perform as well with respect to minimising error. As the ensemble model is considering each of the base model sorts thus taking into account their complementary strength and weakness, the performance of the model improves as compared to individual base models trained, which also accounts for the collective learning of all these base models through this ensemble model approach.

These preprocessing steps also contributed on how well the ensemble performed; that is, outlier removal, balanced data and selecting the best features to guarantee that the input data was of high quality and well formatted for modeling. These steps helped reduce noise and enabled the model better to capture relevant signals. Furthermore, the parameter selection (n estimators, random state) for all models guaranteed the reproducibility and stability of the results.

This performance speaks to the power and generalizability of the stacking ensemble method, which is especially effective for solving complex problems, such as predicting real estate prices, where many factors impact the result in a non-linear way. Also the great R^2 score confirms the model generalizes well to new data. The ensemble model has a tradeoff that it provides a better predictive performance than the other models but results in less of a model fitting and interpretability to the data generated by the models since

the ensemble model can be complex and will have a higher cost of training compared to the individual models around which it is built. Future work might explore the trade-off from a different angle, assessing various meta-models or hyperparameter-search to find an optimal combination.

The ensemble model consistently outperforms all the independent base models and thus represent a solid methodology to approach to the challenges of complex datasets with accurate and reliable predictions. This makes it an effective tool for stakeholders that operate in the field of real estate, as it supports data driven decision making and market analytics.

Study did not perform the cross validation of accuracy due to the complexity of ensembled model with the combination of different base model like Random Forest, XGBoost, LightGBM and the large amount of data might have created hardware limitation and it would have been time consuming for this reason study didn't performed any type of cross-validation

7 Conclusion and Future Work

The specific research question of this study was "How effective staking ensemble predictive model approach for forecasting the Irish real estate price?." Specifically, the study aims to develop a more accurate predictive model for real estate price estimates, tackle data preprocessing issues, assess the impact of data balancing on model performance, and examine the evidence of a real estate bubble in particular counties in Ireland. The study used a stacking ensemble with Random Forest, XGBoost, and LightGBM as base-models, and a Random Forest meta-model to this end. All above preprocessing steps such as outlier removal, feature selection and data balancing helped in ensuring that the dataset is well-prepared for the analysis.

Leveraging a strong model, we could meet one of the aims of the study, evidenced by a MSE of 941.89 and an R^2 score of 0.9999999449 for the ensemble model. This model shows an application of accurate and reliable predictions for real estate prices. Year-on-year growth trends show that property prices and average salaries in both Dublin and Kildare continue to remain healthy, and do not indicate a property bubble. Rather, the trends indicated that observed disparities were explained by imbalances between supply and demand.

7.1 Futurework

Further tuning of hypersearched hyperparameters for base and meta-models may be explored in future work to optimize the stacking ensemble method. Further, by integrating time-series models for capturing temporal trends better and socio-economic factors such as inflation and employment rates; we can add to accuracy of the prediction. The study's framework can also be applicable to commercial use cases like real estate valuation tools, market analysis platforms, and policy development support systems, facilitating data-driven decision-making for stakeholders Alshammari (2023a).

Overall, this paper offers an advanced and comprehensive solution for predicting price in real estate by examining the key challenges of preparing the data to feed to the model, and the actual selection of the model. It not only provides accurate results but also gives a deeper understanding of the market, which could be useful for future research and commercial solutions.but there are few things which it may fall behind the model

price prediction is dependent on steady growth of real-estate if there any kind of natural or man made disaster like earthquake, tsunami, pandemic, riot etc this will effected the real-estate prices but the prediction model is not trained on Realtime data accuracy will decrease's because the model will give output based on old trained data that's why as for future work real time data training with socio economic factors with effect price on grand scale can help the model to predict the property price with best accuracy in dynamic market with its trend and maintain that accuracy for long term.

References

Agnew, K. and Lyons, R. C. (2018). The impact of employment on housing prices: Detailed evidence from fdi in ireland, *Regional Science and Urban Economics* **70**: 174–189.
URL: *https://www.sciencedirect.com/science/article/pii/S0166046217303897*

Alshammari, T. (2023a). Evaluating machine learning algorithms for predicting house prices in saudi arabia, *2023 International Conference on Smart Computing and Application (ICSCA)*, pp. 1–5.

Alshammari, T. (2023b). Evaluating machine learning algorithms for predicting house prices in saudi arabia, *2023 International Conference on Smart Computing and Application (ICSCA)*, pp. 1–5.

Corrigan, E., Foley, D., McQuinn, K., O'Toole, C. and Slaymaker, R. (2019a). Exploring Affordability in the Irish Housing Market, *The Economic and Social Review* **50**(1): 119–157.

Corrigan, E., Foley, D., McQuinn, K., O'Toole, C. and Slaymaker, R. (2019b). Exploring Affordability in the Irish Housing Market, *The Economic and Social Review* **50**(1): 119–157.

Lima, V. (2019). Towards an understanding of the regional impact of airbnb in ireland, *Regional Studies, Regional Science* **6**(1): 78–91.

Limsombunchai, V., Gan, C. and Lee, M. (2004). House price prediction: Hedonic price model vs. artificial neural network, **1**: 193–201.
URL: *https://thescipub.com/abstract/ajassp.2004.193.201*

Madhuri, C. R., Anuradha, G. and Pujitha, M. V. (2019a). House price prediction using regression techniques: A comparative study, *2019 International conference on smart structures and systems (ICSSS)*, IEEE, pp. 1–5.

Madhuri, C. R., Anuradha, G. and Pujitha, M. V. (2019b). House price prediction using regression techniques: A comparative study, *2019 International Conference on Smart Structures and Systems (ICSSS)*, pp. 1–5.

Moro, M., Mayor, K., Lyons, S. and Tol, R. (2011). Does the housing market reflect cultural heritage? a case study of greater dublin, *Environment and Planning A* **45**.

Peng, Z., Huang, Q. and Han, Y. (2019). Model research on forecast of second-hand house price in chengdu based on xgboost algorithm, *2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT)*, pp. 168–172.

Rana, V., Mondal, J., Sharma, A. and Kashyap, I. (2020). House price prediction using optimal regression techniques, pp. 203–208.

Singh, B., Singh, J., Kumar, A. and Gupta, S. (2023). Investigating the effectiveness of machine learning techniques for real estate price prediction, *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*, pp. 1684–1688.

Soibelman, L. and González, M. (2002). A knowledge discovery in databases framework for property valuation, *Journal of Property Tax Assessment and Administration* **7**: 77–106.

Wandhe, A., Sehgal, L., Sumra, H., Choudhary, A. and Dhone, M. (2023). Real estate prediction system using ml, *2023 11th International Conference on Emerging Trends in Engineering Technology - Signal and Information Processing (ICETET - SIP)*, pp. 1–4.