A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Felipe Luiz Bauer Steudel |
| *Student Number* | 2024153 |
| *Module Title* | Machine Learning |
| *Assessment Title* | CA1 Project |
| *Assessment Due Date* | 21st April 2024 |
| *Date of Submission* | 21st April 2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[Importance of Analysis 3](#_Toc164626258)

[Problems to be Addressed 3](#_Toc164626259)

[Future Price Determination 3](#_Toc164626260)

[Justification for Using Forecasting and Classification Algorithms 3](#_Toc164626261)

[Project Objectives 3](#_Toc164626262)

[Characterization 4](#_Toc164626263)

[About Dataset 4](#_Toc164626264)

[Context 4](#_Toc164626265)

[Content 4](#_Toc164626266)

[Acknowledgements 4](#_Toc164626267)

[Data Preprocessing 4](#_Toc164626268)

[Data Cleaning 4](#_Toc164626269)

[Data Preparation 5](#_Toc164626270)

[Conversion of Categorized Values to Numeric 5](#_Toc164626271)

[Hyperparameter Tuning 5](#_Toc164626272)

[GridSearchCV Results 5](#_Toc164626273)

[Model Selection and Interpretation of Results 5](#_Toc164626274)

[Linear Regression 6](#_Toc164626275)

[Advantages and Disadvantages 6](#_Toc164626276)

[Interpretation of results 6](#_Toc164626277)

[Random Forest 6](#_Toc164626278)

[Advantages and disadvantages 6](#_Toc164626279)

[Interpretation of results 7](#_Toc164626280)

[Conclusion 7](#_Toc164626281)

[Recommendations and Improvements 7](#_Toc164626282)

[References 8](#_Toc164626283)

Housing Price Analysis in Ireland

Real estate has traditionally been one of the most important assets in any economy. Inflation in the market is always a good indicator of a country’s economy. The property market in Ireland, as in many other European countries, is driven by political, economic and social factors, and as a result, has seen ups and downs over the years.

# Importance of Analysis

The real estate market is huge, complex and data rich. Humans can only make ambiguous inferences based on general patterns; This is why conventional analytical methods are largely incomplete in accuracy and comprehensiveness. Machine learning algorithms can address this limitation, and have the means to consider large amounts of data and then accurately identify classification and prediction patterns.

This is important because it helps investors and consumers understand price trends; It involves buying, selling and investing in the farm. The methods enable politicians and city planners to elaborate on appropriate housing policies based on market conditions in specific areas.

# Problems to be Addressed

## Future Price Determination

Negotiating effective asset prices is critical for investors and buyers to plan appropriately. Predictive systems can identify price movements based on historical data, reduce market uncertainty, and create better strategies in the real estate market.

# Justification for Using Forecasting and Classification Algorithms

Machine learning algorithms can identify complex patterns in data that often go unnoticed by the most traditional analytical approaches. As for housing prices:

1. Forecasting algorithms can model relationships among multiple variables and predict future values ​​more accurately.
2. Classification algorithms help classify data into categories or groups, such as price categories or market categories (market price vs. non-market price, for example).

# Project Objectives

Predict future property prices in Ireland, using historical data to train predictive models. This objective aims to provide valuable insights that can benefit not only investors and planners but also ordinary individuals in their real estate decision-making processes.

# Characterization

## About Dataset

### Context

All residential properties sold in Ireland from 2010 to May 28th, 2021.

Dataset retrieved from Kaggle at the address: <https://www.kaggle.com/datasets/erinkhoo/property-price-register-ireland/data>

Collaborator: Erin Khoo (Owner)

### Content

**Rows**: 476,745

**Columns**: 9

**Memory usage**: 32.7MB

**Variables**:

* SALE\_DATE => Date of sale (dd/MM/yyyy) | datetime64[ns]
* ADDRESS => Address | string
* POSTAL\_CODE => Postal Code | string
* COUNTY => County | string
* SALE\_PRICE => Price (€) | float32
* IF\_MARKET\_PRICE => Not Full Market Price | int8
* IF\_VAT\_EXCLUDED => VAT Exclusive | int8
* PROPERTY\_DESC => Description of Property | string
* PROPERTY\_SIZE\_DESC => Property Size Description | string

### Acknowledgements

Data sourced from the publicly available site: <https://propertypriceregister.ie>

Source: National Property Price Registry

License: EU ODP Legal Notice

# Data Preprocessing

## Data Cleaning

The following actions were performed to prepare the dataset for analysis.

1. The columns POSTAL\_CODE and PROPERTY\_SIZE\_DESC were removed because they had the highest percentage of null data (81.17% and 88.92%, respectively).
2. Data with invisible characters in the PROPERTY\_DESC column was changed to null and later deleted.
3. 770 records were deleted because they were considered duplicates.

This processing was sufficient to clean the data, remove redundant data and maintain quality.

## Data Preparation

The variable SALE\_DATE was split into SALE\_MONTH and SALE\_YEAR to make it easier to analyze over time, and to allow us to see over time or to group the data by a specific time period.

## Conversion of Categorized Values to Numeric

Converting categorical variables to numeric values ​​allows a wide range of algorithms to be used in machine learning models, and some models require all input variables to be numeric Furthermore, it helps standardize data and simplifies interpretation and analysis.

1. **One-Hot Encoding (Dummy Variables)** - The COUNTY variable contains multiple groups, and the 'dummy variables' method is used to prevent the numeric values ​​assigned to the groups from being interpreted as ordinal values ​​(i.e. , values ​​with a particular order).
2. **Value Variables** - The PROPERTY\_DESC variable has only two values ​​('Second-Hand Dwelling house /Apartment' and 'New Dwelling house /Apartment'), and we can replace them with 0 and 1 respectively.

# Hyperparameter Tuning

Hyperparameter tuning is important for reducing machine learning models. Hyperparameters, which are the settings that determine the training order of an algorithm, must be configured to optimize model performance. Unlike ideal parameters learned during training, hyperparameters are predefined and their variation can significantly affect the results.

Manual tuning of any hyperparameter is usually inefficient and ineffective. Instead, It use structured methods such as GridSearchCV. This method examines a set of parameters identified for each hyperparameter by cross-validation, identifying the combination that gives the best performance on average.

In the linear regression model, it was included K-fold cross-validation and GridSearchCV. By dividing the data into five subsets, the K-fold helps to examine the generalizability of the model to different data segments, and it is likely that it may overfit.

## GridSearchCV Results

**Best Hyperparameters**: {'fit\_intercept': False}

**Best Average Score**: 0.021608914686290514

Setting the optimal setting to 'False' with 'fit\_intercept' results in a low score of about 0.0216, indicating that the linear regression model may not be a good fit for this data set This result have them consider alternative models or strategies for improving efficiency.

# Model Selection and Interpretation of Results

The models chosen for the proposed analysis were linear regression models and random forest models.

## Linear Regression

A linear regression model is a statistical technique that attempts to model the relationship between a dependent variable and one or more independent variables using a straight line Linear regression is a relatively simple model that is easy to interpret and make predictions.

### Advantages and Disadvantages

The model is simple and easy to explain. Training and prediction using linear models is usually fast, especially on small data or with limited features. However, if there is no linear relationship between the features and the target variable, the Linear Regression results may be distorted. If two or more independent variables are highly correlated, this model may not adequately capture the nonlinear relationship between the attributes and the objective variable and in addition is sensitive to outliers ho, which may affect the results of the model.

### Interpretation of results

The presented results show the performance of the linear regression model based on a test data size of 25%.

* **Final R2 Score**: 0.020185897934124286
* **CV Mean**: 0.027273068874712036
* **STD**: 0.012540281502774801
* **Mean Absolute Error**: 135066.27217907988
* **Mean Squared Error**: 740614902523.5021
* **Root Mean Squared Error**: 860589.8573208391

The last hypothesis (R2) indicates that the linear regression model may not perform well on the relationship between the independent and dependent variables in the data the cross-validation mean score (CV Mean) is low, indicating that the model may not such is very strong and may suffer from underfitting. Standard deviation (STD) indicates that the model is sensitive to data distribution and may be difficult for generalizing to new data sets.

The mean absolute error (MAE) is slightly higher, indicating that the model may not adequately capture the relationship between the variables. The mean squared error (MSE) is also very high, indicating that the model does not fit the data well. The root mean square error (RMSE) has a slightly higher value, indicating that the model is not generalizing well on the new data.

In view of these results, the linear regression model is experiencing inappropriate problems. This could be due to the simplicity of the model or the inability of the selected variables to explain the variability in the data.

## Random Forest

Random Forest is an Ensemble machine learning model that uses multiple decision trees for prediction. Random forests combine the predictions of multiple decision trees to yield more accurate and robust information than can be provided by a single decision tree.

### Advantages and disadvantages

One advantage of random forests is that they can be used for backtracking and segmentation. It is considered to be a simple and flexible algorithm because its default hyperparameters tend to yield good predictions. However, one of the biggest problems with Random Forest in terms of machine learning is that it will be too constrained, but if there are enough trees, it won’t happen so easily. However, too many trees can make the algorithm slow and inefficient. More accurate predictions require more trees, which slows down the model.

### Interpretation of results

The presented results show the efficiency of random forest models based on optimized hyperparameters.

* **Best hyperparameters**: min\_samples\_split: 5, n\_estimators: 100
* **Best mean score**: 735508426609.5494
* **Mean Squared Error on Test Set**: 649473049886.5579
* **Coefficient of Determination (R²):** 0.047267925442816106

The optimized hyperparameters for this model have been adjusted to split a minimum of 5 models per node and use 100 decision trees in its forest The mean good score is high, which is typical indicates that the overall model performs well. The errors in the test data are much smaller than the errors in the training set, which may indicate overfitting. A coefficient of determination of 0.049 indicates that the model explains only a small fraction of the variation in the test data, suggesting that the model may not be effective in generalizing to new data

Considering these results, it is possible that the random forest model has a degraded performance, but the reliability may suffer from overfitting due to the large difference in error between the training data and the test data between, and the minimum value of the coefficient of determination.

# Conclusion

The linear regression model showed poor performance, with low R2 and CV Mean. Error metrics (Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error) indicate that the model does not adequately fit the data and makes incorrect predictions

Limitations of this model include its simplicity and the assumption of a linear relationship between the independent and dependent variables, which may not be appropriate for more complex data

On the other hand, the random forest model showed slightly better performance compared to linear regression, with higher R2 and slightly higher CV Mean but still there was room for improvement, as indicated by the error metrics; which is still quite large.

The advantages of this model are its ability to handle nonlinear relationships among variables and its robustness against overfitting due to the combination of multiple decision trees.

## Recommendations and Improvements

In the Linear Regression Model, existing variables can be modified to better capture their complexity data, and outlier treatment not done in preprocessing of the data, but I want to acknowledge that this model is not appropriate for the data set I propose to analyze, because the target variable data do not show linearity with the independent variables.

As for the random forest model, we can rearrange the model’s parameters, such as the number of trees in the forest (n\_estimators) and the split parameters (min\_samples\_split), using the GridSearchCV method to find additional model’s parameters. Another suggestion would be to explore how the size of the data can be reduced to avoid overfitting.

# References

Costa da Silva, J., 2018, Aprendendo em uma Floresta Aleatória (Learning in a Random Forest), <https://medium.com/machina-sapiens/o-algoritmo-da-floresta-aleat%C3%B3ria-3545f6babdf8>

Hericlins, S., 2022, Hiperparâmetros - por quê são importantes [Hyperparameters - why are they important], <https://blog.dsbrigade.com/hiperparametros-por-que-sao-importantes/>

Iqbal, M., Overfitting, Underfitting and Generalization - Week 2. Study Material, CCT College Dublin, Dublin.

Iqbal, M., Decision Trees and Random Forest - Week 3. Study Material, CCT College Dublin, Dublin.

Iqbal, M., Linear Regression - Week 4. Study Material, CCT College Dublin, Dublin.

Koushik, 2023, What is Overfitting and Underfitting , and how to deal with it step by step?, <https://medium.com/@koushikkushal95/what-is-overfitting-and-underfitting-and-how-to-deal-with-it-step-by-step-d6c335ed086b>

Piubello, H., 2023, Regressão Linear: O que é, Modelagem, Avaliação e Interpretação [Linear Regression: What it is, Modeling, Evaluation, and Interpretation], [https://codecrush.com.br/blog/regressao-linear#considera%C3%A7%C3%B5es-finais](https://codecrush.com.br/blog/regressao-linear%23considera%C3%A7%C3%B5es-finais)