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](#_Toc164389963)

[Problems to be Addressed 3](#_Toc164389964)

[Future Price Determination 3](#_Toc164389965)

[Justification for Using Forecasting and Classification Algorithms 3](#_Toc164389966)

[Project Objectives 3](#_Toc164389967)

[Characterization 4](#_Toc164389968)

[About Dataset 4](#_Toc164389969)

[Context 4](#_Toc164389970)

[Content 4](#_Toc164389971)

[Acknowledgements 4](#_Toc164389972)

[Data Preprocessing 4](#_Toc164389973)

[Data Cleaning 4](#_Toc164389974)

[Data Preparation 5](#_Toc164389975)

[Conversion of Categorized Values to Numeric 5](#_Toc164389976)

[Hyperparameter Tuning 5](#_Toc164389977)

[GridSearchCV Results 5](#_Toc164389978)

[Model Selection and Interpretation of Results 6](#_Toc164389979)

[Linear Regression 6](#_Toc164389980)

[Advantages and Disadvantages 6](#_Toc164389981)

[Interpretation of Results 6](#_Toc164389982)

[Random Forest 7](#_Toc164389983)

[Advantages and Disadvantages 7](#_Toc164389984)

[Interpretation of Results 7](#_Toc164389985)

[Conclusion 7](#_Toc164389986)

[Recommendations and Improvements 8](#_Toc164389987)

[References 8](#_Toc164389988)

Housing Price Analysis in Ireland

The real estate sector has always been an important part of the economy, with property prices being a strong indicator of a country's economic condition. In Ireland, as in many other countries, the real estate market has been influenced by political, economic, and social factors that have caused property prices to fluctuate over time.

# Importance of Analysis

The real estate market is complex and vast, containing a huge amount of data. Often, traditional methods of accurately analyzing the various market patterns are flawed and insufficient. Machine learning can bridge this gap, as it has tools capable of processing large volumes of data and making accurate predictions and classifications.

This is crucial for helping investors and buyers understand future property price trends, whether for buying and selling or for making investments in the sector. Machine learning methods can also assist politicians and urban planners in creating better housing policies through detailed analysis of regional market conditions.

# Problems to be Addressed

## Future Price Determination

To adequately plan for investors and consumers, effective property price forecasting is crucial. Forecasting algorithms can identify trends of rising or falling prices based on historical data, reducing market uncertainties and developing better strategies in the real estate market.

# Justification for Using Forecasting and Classification Algorithms

Machine learning algorithms can identify complex patterns in data that are often not visible through more traditional analysis methods. Regarding housing prices:

1. Forecasting algorithms can model relationships between multiple variables and effectively predict future prices.
2. Classification algorithms help categorize data into groups or classes, such as price ranges or market categories (market price vs. non-market, 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

To prepare the dataset for analysis, the following actions were taken:

1. Columns POSTAL\_CODE and PROPERTY\_SIZE\_DESC were removed due to containing a very high percentage of null data (81.17% and 88.92%, respectively).
2. Data with unrecognized characters in the PROPERTY\_DESC column were changed to null and subsequently deleted.
3. 770 records were deleted due to being considered duplicates.

These actions were sufficient for data cleaning, removal of null data, and maintaining them in good condition.

## Data Preparation

The SALE\_DATE variable was split into SALE\_MONTH and SALE\_YEAR to facilitate and simplify temporal analysis, allowing for the identification of seasonal trends over time or grouping data by specific time intervals.

## Conversion of Categorized Values to Numeric

Converting categorical variables into numerical values allows for a wider range of algorithms to be used in the Machine Learning model, and in some models, it is required that all input variables be numeric. Additionally, it helps to standardize the data and make it easier to interpret and analyze.

1. **One-Hot Encoding (Dummy Variables) -** In the COUNTY variable, there are several categories, and the technique of 'dummy variables' is used to ensure that the numerical values assigned to the categories are not interpreted as ordinal values (i.e., values that have a specific order).
2. **Value Transformation** - The variable PROPERTY\_DESC 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 crucial for refining a Machine Learning model. Hyperparameters, which are settings that govern the training process of an algorithm, must be optimally set to enhance model performance. Unlike model parameters that are learned during training, hyperparameters are predefined and their adjustment can significantly impact outcomes.

Commonly, manual tuning of each hyperparameter is inefficient and ineffective. Instead, systematic methods like GridSearchCV are used. This technique explores a range of specified values for each hyperparameter using cross-validation, identifying the combination that yields the best average performance.

In our linear regression model, we integrated K-fold cross-validation with GridSearchCV. By partitioning the data into five subsets, K-fold helps evaluate the model's generalization capability across different data segments, highlighting potential overfitting.

## GridSearchCV Results

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

**Best Average Score**: 0.021608914686290514

The optimal setting with 'fit\_intercept' set to 'False' suggests a low average score of approximately 0.0216, indicating that the linear regression model might not be well-suited for this dataset. This result prompts consideration of alternative models or techniques to improve performance.

# Model Selection and Interpretation of Results

The selected models for the proposed analysis were Linear Regression and Random Forest models.

## Linear Regression

The Linear Regression model is a statistical technique that attempts to model the relationship between a dependent variable and one or more independent variables through a straight line. Linear Regression is a relatively simple model that is easy to interpret and generate predictions.

### Advantages and Disadvantages

It is a simple model and easy to interpret. Training and making predictions with linear models is usually fast, especially on small datasets or with few features. However, Linear Regression results can be distorted if there is no linear relationship between the features and the target variable. This model may not capture nonlinear relationships between the features and the target variable effectively when two or more independent variables are highly correlated. Additionally, it is sensitive to outliers, which can influence the model's results.

### Interpretation of Results

The results provided indicate the performance of the Linear Regression model based on a 25% test data size.

* **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 final coefficient of determination (R²) suggests that the linear regression model may not be very effective at capturing the relationship between the independent and dependent variables in the data. The cross-validation mean score (CV Mean) is low, suggesting that the model may not be very robust and may be suffering from underfitting. The standard deviation (STD) indicates that the model is sensitive to the distribution of the data and may have difficulty generalizing to new datasets.

The mean absolute error (MAE) is relatively high, suggesting that the model may not be adequately capturing the relationship between the variables. The mean squared error (MSE) is also quite high, indicating that the model may not be fitting the data well. The root mean squared error (RMSE) is a relatively high value, suggesting that the model may not be generalizing well to new data.

Considering these results, the Linear Regression model is facing underfitting problems. This may be due to the simplicity of the model or the inadequacy of the chosen variables to explain the variability in the data.

## Random Forest

Random Forest is an Ensemble machine learning model that uses many different decision trees to make predictions. Random Forest combines the predictions of many different decision trees to arrive at a more accurate and robust conclusion than a single decision tree could offer.

### Advantages and Disadvantages

One advantage of Random Forest is that it can be used for both regression and classification. It is considered an easy and accessible algorithm because its default hyperparameters usually produce a good prediction result. However, one of the big problems with Random Forest in machine learning is overfitting, but it won't occur so easily if there are enough trees. However, if there are too many trees, it can make the algorithm slow and inefficient. More accurate prediction requires more trees, which makes the model slower.

### Interpretation of Results

The results provided indicate the performance of the Random Forest model based on the 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 were adjusted to have a minimum sample split at each node of 5 and use 100 decision trees in its forest. The best mean score is high, which usually indicates better overall model performance. The mean squared error on the test data is significantly lower than the error on the training set, which may indicate overfitting. The coefficient of determination of 0.049 indicates that the model explains only a small part of the variability in the test data, suggesting that the model may not be very effective at generalizing to new data.

Considering these results, it is possible that the Random Forest model has moderate performance, but it may be suffering from overfitting due to the significant difference between the error on the training data and the test data, as well as the low value of the coefficient of determination.

# Conclusion

The Linear Regression model showed poor performance, with low R² and CV Mean. The error metrics (Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error) indicate that the model is not fitting well to the data and providing inaccurate predictions.

The limitations of this model include its simplicity and assumption of a linear relationship between the independent and dependent variables, which may not be suitable 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. However, there is still room for improvement, as indicated by the error metrics, which are still quite high.

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

## Recommendations and Improvements

For the Linear Regression Model, transformations can be made to the existing variables to better capture the complexity of the

data, and outlier treatment not performed in the preprocessing of the data, but I am inclined to accept that this model is not suitable for the dataset I proposed to analyze, as the target variable data does not exhibit linearity with the independent variables.

Regarding the Random Forest Model, we can readjust the model's hyperparameters, such as the number of trees in the forest (n\_estimators) and the split criterion (min\_samples\_split), using the GridSearchCV technique to find a new ideal combination. A broader adjustment of the hyperparameters caused a significant delay in training the model, which forced me to reduce the hyperparameters. Another recommendation would be to investigate the possibility of reducing the dimensionality of the data 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)