A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Marty Wickham |
| *Student Number* | sbs24069 |
| *Module Title* | Machine Learning |
| *Assessment Title* | CA1 |
| *Assessment Due Date* | 28/04/2024 |
| *Date of Submission* |  |

**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

[Machine Learning CA1 1](#_Toc165407349)

[Introduction: 1](#_Toc165407350)

[Data Selection 1](#_Toc165407351)

[Data Preprocessing 1](#_Toc165407352)

[Test-Train Split and Cross Validation 2](#_Toc165407353)

[Test-Train Split Sizes 2](#_Toc165407354)

[Test-Train Split Size Results 2](#_Toc165407355)

[Hyperparameter Tuning 4](#_Toc165407356)

[Results Interpretation 5](#_Toc165407357)

[Linear Regression 5](#_Toc165407358)

[kNN Regression 5](#_Toc165407359)

[References 6](#_Toc165407360)

# Machine Learning CA1

## Introduction:

For this project, I have chosen to develop and deploy two machine learning models that will be capable of accurately predicting property prices, based on universal features that are inherent in all properties. Ireland is currently in the middle of a serious housing crisis with house prices becoming increasingly more expensive, with “tens of thousands of young people locked out of the housing market” (Conneely, 2024). Many young people feeling like the dream of owning a home has never been so far out of reach. The overarching goal of this project will be to develop a model that will be capable of providing prospective homeowners with an increased measure of control in their housing purchases. Considering the unpredictable trends driven by external factors such as scarcity, inflation, and energy costs, this tool could provide valuable guidance for buyers and policy makers and an understanding of the valuations in the housing retail market. First-time buyers could benefit from a tool which could give them reasonably accurate estimates based on their individual requirements and preferences, offering insight into local area affordability, empower negotiation exchange and narrowing of the decision-making process based on means ands requirements. This tool could further guide those selling or renovating properties to choose their asking prices and renovation investments based on trends elucidated from a ML market valuation analysis. Predictive modelling techniques implementing prediction and classification algorithms have emerged as indispensable tools for navigating complex domains such as this, especially for the average individual outside of the housing industry. This project seeks to demonstrate the utility of these algorithms to ameliorate these challenges and provide actionable insights in the domain of property valuation and investment decision-making.

## Data Selection

To develop models capable of predicting property prices, I decided to use the “House Sales in King County, USA” dataset. This dataset contains house sale prices for King County, Seattle and includes homes sold between May 2014 and May 2015. My reasons for choosing this data are as follows:

* It is a dynamic dataset which provides a wide variety of property features to analyse.
* It is a very complete dataset that provides feature value for all observations.
* It is an exceptionally large dataset providing ample data on which to train the machine learning models.
* The dataset is well established as a reliable source evaluating for simple regression models ()

## Data Preprocessing

The dataset was imported as a Pandas data-frame object and analysed utilising the classes built in functionality. Firstly, I used the .head(20) method to give me a brief overview of the data, followed by the .describe() method to provide me with a numerical summary the values contained in each feature. The .info() method allowed me to identify any categorical features and any instances of NULL values. The results of this exploratory analysis revealed that the dataset is very clean containing no NULL values and only 1 categorical feature, date. Once I had identified that there were no NULL values to be cleaned or categorical features to be converted, I dropped the unnecessary features ‘id’ and ‘date’ from the data frame. After examining the data further, I noticed that there seemed to be a high number of 0 values in the column ‘yr\_renovated’. I discovered that there were 20699 values of 0, so I decided to remove this column from the dataset as well.

## Test-Train Split and Cross Validation

### Test-Train Split Sizes

Test-training split sizes play an important role in development and evaluation of machine learning models. Depending on the type of model being used it is important to experiment with different values for the training and testing split size, to ensure that the model can achieve the best performance possible. For instance, in some cases a lager training set can be important as it can provide the model with the required data to learn from effectively and achieve the desired performance and generalization. On the other hand, smaller training sets may result in underfitting, where the model has not learned well from the training data and is unable to acquire the underlying patterns in the training data resulting in high bias and poor performance. Overall, the most optimal test-training split used will depend on the specific nature of the dataset and the model deployed. In this project I have decide to test my algorithms over 4 different training splits of 15, 20, 25 and 30 percent. From my understanding, this should be a robust enough variation to ensure that my models can achieve optimal performance from a test-training split point of view.

### Test-Train Split Size Results

#### Linear Regression:

A screenshot of a computer code

Description automatically generated A screenshot of a white screen

Description automatically generated

#### kNN Regression:

A screenshot of a computer program

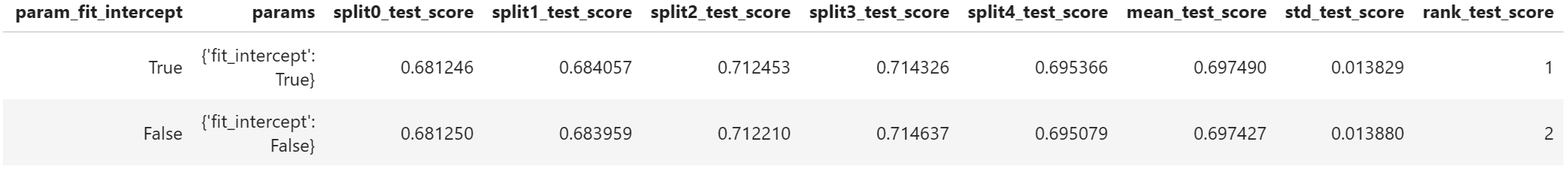
Description automatically generated

The results of my test-training splits show that the most optimum split values for the linear regression and kNN regression models are 25% and 15% respectively. Results showed that there was not a huge performance increase to be gained by using different splits on this data. I would speculate that this is due to the overall size of the dataset which contains over 21,000 observations and is more than enough data for either algorithm to be trained with.

In addition to this, I have employed cross validation within my model training splits. Cross validation is the concept of creating multiple training splits compromised of variations in the selection of randomly chosen records. In the event where feature representation imbalance plagues the dataset, this can have an impact on model performance, as an imbalanced selection of data for a training split may result in the model learning patterns and trends that do not accurately represent the data.

Cross validation can help to mitigate the consequences of this, by creating variability in the distribution of observations within training subsets. For my linear regression model, I have opted to choose 8 cross validation splits. I believe this should be more than enough splits to account for any imbalances in the data as the dataset I have chosen is very well recorded and maintained with features being well represented across the dataset. For my kNN regression model I have opted only for 5 splits due to the performance overheads required by the model which result in much slower prediction times.

#### Linear regression cv splits:



The results of my splits in the linear regression model show a variation in accuracy between 68 – 71% which are a good indication that dataset used is well balanced.

## Hyperparameter Tuning

In machine learning models, hyperparameters are parameters which “express important properties of the model such as its complexity or how fast it should learn”. Hyperparameters can be thought of as configuration settings for a model which are set before the model is trained. They are ultimately responsible for controlling how the model learns from the data it is trained with. Some of examples this are the number of neighbours to search in a kNN regression, the kernel size in a support vector machine or the number of neurons in a layer of a neural network. Hyperparameter tuning, is the process of adjusting these configuration variables to identify the optimal hyperparameter values which will result in the best performance for a model (GeeksForGeeks, 2019).

For this project I have decided to use GridSearchCV as the method for controlling hyper parameter tuning. GridSearchCV is a hyperparameter tuning approach that allows a model developer to provide a specified list of relevant hyperparameters and values to be employed during model training. Using the specified list, GridSearchCV will fit the model using all possible combinations of hyperparameters in a systematic brute force approach. Each performance outcome of the model is recorded, and the model with the best performance will be chosen as the hyperparameters to be used.” The approach is called GridSearchCV, because it searches for the best set of hyperparameters from a grid of hyperparameters values” (GeeksForGeeks, 2019). Aside from the key advantage of producing the best accuracy for a model, another key advantage of the approach is that it naturally incorporates cross validation splits discussed previously, with a default value of 5 cv splits if another is not specified. Although this approach will identify the best performance for a model, a primary disadvantage of the approach is the time and space complexity of the search method. As an exhaustive approach is employed this increase the computational overhead required to run the search, which may be a limiting factor. As an outcome of this, devices with less computational power may result in longer times for results.

# Results Interpretation

### Linear Regression

R-squared (R²) Testing: 0.567

The R-squared value indicates that 57% of the variability in the target variable is explained by the model. This suggests that the model is not performing very well in predicting the target variable on unseen data.

R-squared (R²) Training: 0.537

The R-squared value indicates that 54% of the variability in the target variable is explained by the model during training. This suggests that the model does captures a substantial portion of the variability in the training data and may exhibit some degree of underfitting.

Mean Absolute Error (MAE): 153303.001

The MAE measures the average absolute difference between the predicted and actual values of the target variable.

An MAE of 153303.001 suggests that on average, the model's predictions are off by approximately 153303.001 units.

Root Mean Squared Error (RMSE): 257091.300

The RMSE measures the square root of the average squared difference between the predicted and actual values of the target variable.

An RMSE of 257091.300 suggests that on average, the model's predictions are off by approximately 257091.300 units.

Overall, these results suggest that the linear regression model provides a decent fit to the data and indicates that the model generalizes well to unseen data. If training R-squared was much higher than the test R-squared scores, this would suggest overfitting. If both scores were low, this would suggest underfitting. While the R-squared value of the model was relatively good, MAE and the RMSE are very high at 126863.168 and 210130.343 respectively. This indicates that the models’ predictions are not as accurate as the R-squared value would suggest. These results could possibly be explained by a significant number of outliers in the data. For example, the maximum price value in the dataset is 7,700,000 which is extremely above the mean value of

### kNN Regression

R-squared Testing: 0.567

The R-squared value indicates that 57% of the variability in the target variable is explained by the model. This suggests that the model is not performing very well in predicting the target variable on unseen data.

R-squared Training: 0.537

The R-squared value indicates that 54% of the variability in the target variable is explained by the model during training. This suggests that the model does captures a substantial portion of the variability in the training data and may exhibit some degree of underfitting.

Mean Absolute Error: 153303.001

An MAE of 153303.001 suggests that on average, the model's predictions are off by approximately 153303.001 units.

Root Mean Squared Error: 257091.300

An RMSE of 257091.300 suggests that on average, the model's predictions are off by approximately 257091.300 units.

Since RMSE penalizes large errors more than MAE, the higher RMSE indicates that the model's predictions may have larger deviations from the actual values.

Overall, these results indicate that the kNN regression model only explains a moderate amount of variability in the target variable and only does mediocre with predicting unseen data. There is room for improvement, particularly in reducing prediction errors and potentially addressing underfitting in the training data. Further optimization or feature selection may be warranted to enhance performance.

# References