A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name:* | Kevin Scully |
| *Student Number* | 24275 |
| *Module Title* | Machine Learning |
| *Assessment Title* | CA1 |
| *Assessment Due Date* | 24th November 2024 |
| *Date of Submission* | 18th November 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

[Introduction 3](#_Toc182866184)

[Data Cleansing 3](#_Toc182866185)

[Data Pre-Processing 3](#_Toc182866186)

[Model Building 4](#_Toc182866187)

[Cross Validation 5](#_Toc182866188)

[Hyperparameter tuning 5](#_Toc182866189)

[Conclusion 6](#_Toc182866190)

[References 6](#_Toc182866191)

[Appendix: 8](#_Toc182866192)

# Introduction

A current major issue for the Irish government is the shortfall in supply of social housing to meet the ever-increasing demand. The lack of social housing has a direct impact on homelessness figures, increasing the number of families living on the street. (Clarke, et al., 2024) Therefore, social housing is a key pillar of current economic policy.

This report will aim to help the government departments make decisions on the construction of social housing by using machine learning models to analyse the following objective:

1. Build a machine learning model that best predicts how many social houses will be constructed based on the information provided within the dataset.

The dataset used for this analysis was the ‘Social Housing Construction Status Report Q2 2022’. This provided data on the construction of social homes within Q2 2022. The information within the data includes the funding programme, the local authority, the scheme/project name, the number of units constructed, the approved housing body, and whether the construction was on-site or completed. (Department of Housing, Local Government and Heritage, 2022)

# Data Cleansing

The dataset was analysed to understand the columns and values. There was 1,566 rows and 12 columns within the dataset. After analysing the missing values within each column, I found that the columns “Stage 1 Capital Appraisal”, “Stage 2 Pre-Planning”, “Stage 3 Pre-Tender design” and “Stage 4 Tender Report or Final Turnkey/CALF approval” had over 78% of the rows as null values in each column. These columns did not provide enough information to be analysed and were removed.

The columns “On Site” and “Completed” also had a significant number of missing values. A missing value here was understood to mean that the construction was not on site or completed. Therefore, these null values were replaced with “No”. Finally, the column “Approved Housing Body” had 42 missing values, which equated to 2.7% of all the rows. As this was an insignificant number of rows, these were dropped from the dataset.

# Data Pre-Processing

Now that the data was clean, it was ready for pre-processing. The data contained a mix of categorical and numerical data types. The label encoder was used to convert the categorical data types into numerical data types. When this was complete, a target variable was defined. As the objective of this analysis was to predict the features that would impact the number of social houses that got constructed, this was picked as the target variable, with all other columns therefore being defined as the feature variables.

The target variable was converted to a categorical data type. As there was varied data as values within the target variable, these were classified into two ranges, 0-11 houses constructed and 12+ houses constructed.

# Model Building

The target variable was tested across five different machine models using a test split of 10%, 15% and 25%. The five models were k-nearest neighbour, decision tree, random forest, logistic regression and support vector machine. The results were as follows:

A screenshot of a computer

Description automatically generated

As the changes in accuracy across the training sets did not vary greatly (variance between 0% and 3%), this shows that this is a stable model that is likely to generalise well to new subsets of data. It also shows that the data is well balanced. (Rajput, et al., 2023) The best performing models are the decision tree and random forest, scoring at 67% and 68% respectively using the 75% training set. This shows that these models perform moderately at making predictions.

# Cross Validation

A 5-fold cross-validation was used due to computation power required to complete the 10-fold cross-validation. As the data size was 1,524 rows, the 5-fold was an acceptable measurement size. The results were as follows:

A white background with black text

Description automatically generated

The results from cross-validation show that the model is generalising better with the full dataset. The small decrease in accuracy shows that the model is learning the general patterns and not memorising specific details from the training data. This is normal behaviour following cross validation and shows a more accurate performance of the model. (Japkowicz &  Shah, 2011) Following cross-validation the logistic regression was best performing models. The decision tree and random forest were still performing moderately.

# Hyperparameter tuning

Hyperparameter tuning is used to optimise the machine learning models performance by finding the best combination of hyperparameters that lead to the highest percentage of accuracy for each model. (Shivaiah, et al., 2024) The Grid Search CV was used to establish the best hyperparameter for each of the models. This is a technique that complete the hyperparameter tuning by using the cross-validation results to complete an exhaustive search over a defined grid. (Montesinos López, et al., 2022)

The Grid Search CV was applied to 75% training set with the following results:

A screenshot of a test

Description automatically generated

This was then generalised successfully to the full datasets to provide the final results of:

A white background with black text

Description automatically generated

# Conclusion

The objective of this assignment was achieved. Five models were tested to find the best model to predict the number of social houses that were constructed based on the parameters within the dataset. The model was cross validated and the hyperparameter technique of Grid Search CV was used. The cross-validated results shown that there was no imbalance in the data, and the data was generalised successfully. The low impact on accuracy of the generalised data showed that the data was neither overfitted nor underfitted. The best model for predicting the number of social houses that were constructed based on the parameters within the dataset was logistic Regression which had an accuracy of 60% following hyperparameter and cross validation. The code and description of steps is provided as an appendix.

# References

Clarke, A., Cheshire, L., Parsell, C. & Morris, A., 2024. Reified scarcity & the problem space of ‘need’: unpacking Australian social housing policy. *Housing Studies,* 39(2), pp. 565-583.

Department of Housing, Local Government and Heritage, 2022. *Social Housing Construction Status Report Q2 2022.* [Online]   
Available at: https://opendata.housing.gov.ie/dataset/social-housing-construction-status-report-q2-2022  
[Accessed 14 November 2024].

Japkowicz,  . &  Shah,  ., 2011. *Evaluating Learning Algorithms: A Classification Perspective.* United States:: Cambridge University Press.

Montesinos López,  .,  Montesinos López,  . &  Crossa,  ., 2022. *Multivariate Statistical Machine Learning Methods for Genomic Prediction.* Poland: Springer International Publishing.

Rajput, D., Wang, W. & Chen, C., 2023. Evaluation of a decided sample size in machine learning applications. *BMC bioinformatics,* 24(1).

Shivaiah, S. et al., 2024. Prediction of mental illness using ensemble model and grid search hyperparameter optimization.. *Bulletin of Electrical Engineering and Informatics,* 13(5), pp. 3726-2736.

# Appendix:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A group of black text

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated