A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Suzanne Cotter* |  |
| *Sbs24020* |  |
| *Machine Learning* |  |
| *CA1* |  |
| *28th April, 2024* |  |
| *28th 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

[Introduction 1](#_Toc164778229)

[Data Characterization and Preprocessing: 1](#_Toc164778230)

[Chapter 1 3](#_Toc164778231)

[Chapter 1.1 3](#_Toc164778232)

[Chapter 1.1.1. 3](#_Toc164778233)

[Conclusion 4](#_Toc164778234)

[References 4](#_Toc164778235)

## Introduction

Predicting house prices and understanding which housing characteristics are the most important is vital in real estate markets, urban planning and personal finance decision-making.

#### Description of the Problem Domain:

When predicting housing prices, the primary challenge lies in accurately estimating the value of a property based on various features and characteristics. This task is complex due to the many factors influencing housing prices, including location, size, condition economic indicators, and demographic factors. These factors often interact with each other in nonlinear ways, further complicating the prediction process.

Key challenges in predicting housing prices accurately include:

* Data Variability: Housing markets differ a lot from place to place, even within the same city. For example, a house in one neighbourhood might be more expensive than a similar house in another area. Predictive models need to understand these differences while still being useful in different places.
* Feature Selection: Figuring out which details about a house are most important for predicting its price is vital. However, it's not always easy to know which details matter the most or how they work together.
* Data Quality: Making sure that the information we have about houses is accurate and reliable is important for making good predictions. Sometimes, the data we have might be missing or wrong, and that can mess up our predictions if we don't handle it carefully.
* Model Complexity: Predictive models can be simple or complex. Simple models are easy to understand but might miss some important details. Complex models can capture a lot of detail, but they might also make mistakes if they're too complicated.
* Market Dynamics: Housing markets change over time because of things like the economy, government rules, and how people live. Predictive models need to be able to adapt to these changes so they can keep making good predictions.

The Ames Housing dataset is a collection of data about houses in Ames, Iowa, USA.

The dataset consists of 2930 rows, with each row representing a different house, and 82 columns representing different features or attributes of the houses.

It is a perfect subject for this project due to the following reasons:

* It is relevant in the real world as it is about people’s homes and describes features such as size, location and amenities.
* It has a wide range of features, both numerical and categorical which make it perfect for testing and training prediction models on.
* It is large enough and complex enough to showcase various machine learning techniques and offers challenges like missing data, categorical variables and feature engineering opportunities.
* There is a lot of information available on this dataset making it a good start for a first assignment in a Machine learning module.

## Data Characterization and Preprocessing:

Common features in the dataset include the following:

* Numeric features: These include different types of variables like 'Lot Area' (the size of the lot in square feet), 'Year Built' (the year the house was built), 'Bedroom AbvGr' (the number of bedrooms above ground), and 'SalePrice' (the price at which the house was sold).
* Categorical features: These include variables like 'Neighborhood', 'Exterior 1st' (the exterior covering of the house), 'Heating' (the type of heating system), and 'Garage Type' (the type of garage, if any).

It's important to understand the structure of the dataset, including the number of rows and columns, the data types of each column (numeric, categorical, etc.), and the distribution of values within each column.

#### Data Preprocessing Steps:

* Handling Missing Values: Check for missing values in the dataset. We filled in Categorical blanks with NA and numerical blanks with 0.0.
* Encoding Categorical Variables: Convert categorical variables into a numerical format that machine learning algorithms can understand. We used one-hot encoding and .

Scaling Numerical Features: Scale numerical features to ensure that they have a similar range of values. Common scaling techniques include min-max scaling (scaling values to a range between 0 and 1) or standardization (scaling values to have a mean of 0 and a standard deviation of 1).

Explanation of Techniques to Handle Data Anomalies or Outliers:

Identifying Outliers: Outliers are data points that deviate significantly from the rest of the data. They can skew statistical analyses and machine learning models if not handled properly. Techniques for identifying outliers include visual inspection using box plots or scatter plots, or statistical methods such as z-score or interquartile range (IQR) method.

Handling Outliers: Once outliers are identified, they can be treated in various ways, depending on the nature of the data and the analysis being performed. Some common approaches include removing outliers, transforming variables to make the distribution more symmetric, or applying robust statistical methods that are less sensitive to outliers.

By performing these data characterization and preprocessing steps, we can ensure that the dataset is clean, properly formatted, and ready for analysis and modeling. This lays the foundation for building accurate and reliable predictive models for housing prices.

o Which classification approach do you prefer for the prediction of X as a target variable, and why?

o How to classify the loyal and churn customers using Support Vector Machines?

o Why is dimensionality reduction important in machine learning?

The student would need to consider the following instructions (a - d) during the development of this

project.

a) Logical justification based on the reasoning for the specific choice of machine learning approaches.

b) Multiple machine learning models (at least two) using hyperparameters and a comparison between

the chosen modelling approaches.

c) Visualise your comparison of ML modelling outcomes. You may use a statistical approach to argue that

one feature is more important than other features.

d) Cross-validation methods should be used to justify the authenticity of your ML results.

You will present the findings and defend the results in the report (MS Doc) by highlighting your work. Your

report should capture the following aspects that are relevant to your project investigations.

1. A clear introduction, motivation, a description of the problem domain, and an explanation of how the

project's goals are justified using Prediction / Classification algorithms.

(20 marks)

2. Characterization of data, pre-processing, explanation and description of techniques used for the

variation in the accuracy across three training splits (20%, 25% and 30%) using cross validation

techniques.

(30 marks)

3. What is the primary purpose of hyperparameter tuning in machine learning? Could you elaborate on

specific hyperparameter tuning techniques (e.g., GridSearchCV) applied to machine learning models

to find optimal parameters?

(25 marks)

4. Interpret and explain the results obtained, discuss overfitting / underfitting / generalisation, provide a

rationale for the chosen models and use visualisations to support your findings. Comments in Python

code, conclusions of the project should be specified at the end of the report. Harvard Style must be

used for citations and references.

# Chapter 1

## Chapter 1.1

### Chapter 1.1.1.

A screenshot of a computer

Description automatically generated

[NOTE: For the table of contents to function properly, you must use the correct headings for all your chapters and subchapters.

**Heading 1:** This is the main heading and should be employed for the primary title or chapter. For example: CHAPTER 1.

**Heading 2:** Use Heading 2 as a subheading. For instance: Chapter 1.1.

**Heading 3:** Heading 3 provides a more detailed breakdown, such as Chapter 1.1.1.

By adhering to this hierarchical structure, you ensure an organized and effective document outline, enhancing readability and navigation. However, you are not forced to use all 3 headings, usually heading 1 and 2 are sufficient.

The remainder of your text should be written using a normal font.]

# Conclusion

# References

De Cock, D. (2023). *Ames Iowa: Alternative to the Boston Housing Data Set*. [online] Amstat.org. Available at: https://jse.amstat.org/v19n3/decock/DataDocumentation.txt.