# Group ID - MSc in Data Analytics

Author: Sonia Guzman

e-mail: 2023070@student.cct.ie

Student ID: 2023070

## Abstract

*This paper provides an overview of the average price of a new house in Ireland over the period 1997 to 2015. It reviews the factors that drove the fluctuations and the factors that influenced those fluctuations in the main cities of the country during the mentioned period, with a particular focus on Dublin. The analysis is based on the index Average New House Price, a dataset the Department of Housing Local Government and Heritage in the domain of Housing constructs using data supplied by the mortgage lending agencies on loans approved by them rather than loans paid. It is also important to highlight that the values in the dataset are a result of a mix of houses (including apartments). The study will utilize a few statistical techniques and data visualization tools such as to analyse and interpret the data, and we will try to build a model using Linear Regression method of Machine Learning for predicting property prices.*

*Keywords: fluctuations, factors, average, new house price, Dublin*

**Table of Contents**

[Group ID - MSc in Data Analytics 1](#_Toc132310958)

[Abstract 2](#_Toc132310959)

[Introduction 5](#_Toc132310960)

[Materials and Methods 7](#_Toc132310961)

[Project Implementation Tools and Programming Language 7](#_Toc132310962)

[Project Management Methodology: CRISP-DM Model 7](#_Toc132310963)

[Real-life Scenario 9](#_Toc132310964)

[The CRISP-DM Methodology Applied 10](#_Toc132310965)

[Business Understanding 10](#_Toc132310966)

[Data Understanding 10](#_Toc132310967)

[Collect the Data 11](#_Toc132310968)

[EDA and Data Preparation 11](#_Toc132310969)

[Data Exploration/Visualization 14](#_Toc132310970)

[Statistics: Distribution 18](#_Toc132310971)

[Model Building and Results 20](#_Toc132310972)

[Model Building 20](#_Toc132310973)

[Evaluation 23](#_Toc132310974)

[Deployment 23](#_Toc132310975)

[Conclusion 25](#_Toc132310976)

[Acknowledgements 26](#_Toc132310977)

[References 27](#_Toc132310978)

**List of Figures**

[Figure 1: Six Phases of CRISP-DM. Image taken from Project Management Methodologies Week 4 PPT. CCT College Dublin 2022 9](#_Toc132310583)

[Figure 2: First Five Rows of the Dataset after Dropping the original Index 12](#_Toc132310584)

[Figure 3: Field Types of the Dataset 12](#_Toc132310585)

[Figure 4: Updated Dataset 13](#_Toc132310586)

[Figure 5: Structure of the Dataset 13](#_Toc132310587)

[Figure 6: Null Value Detection 14](#_Toc132310588)

[Figure 7: Heatmap of Features Correlation 15](#_Toc132310589)

[Figure 8: Average House Price in Ireland (1997-2015) 16](#_Toc132310590)

[Figure 9: Percentage of difference between one year and the next 17](#_Toc132310591)

[Figure 10: Box plot of Mean House Prices by City 17](#_Toc132310592)

[Figure 11: Box plot of Mean House Prices by City 18](#_Toc132310593)

[Figure 12: KDE Plot of Dublin 19](#_Toc132310594)

[Figure 13: Calculates the probability that a new house in Dublin with a price of €400000 or less will be sold 19](#_Toc132310595)

[Figure 14: Calculates the probability that a new house in Dublin being sold for more than €400000 20](#_Toc132310596)

[Figure 15: Machine Learning 22](#_Toc132310597)

[Figure 16: GridSearchCV 22](#_Toc132310598)

## Introduction

The real estate market can be a good indicator of the health of the state’s economy, and

We have used python Jupiter Notebook for a simple visualization of the used dataset.

## 

## Datasets

## General Information About the Dataset – Country: Ireland.

Sector: Construction.

Topic: New Dwelling Completions.

Period: 2011 – 2022.

Source: New Dwelling Completions, [Central Statistics Office](https://data.gov.ie/organization/central-statistics-office) (CSO). Link: <https://data.gov.ie/dataset/nda02-new-dwelling-completions>

Licensed under: Creative Commons Attribution 4.0.

Data set name file: NDA02.20230515111519.csv

Data Structure: The data is structured in a time-series format and each observation is recorded annually at a national level.

Variables: The original dataset is composed by eight (8) variables, as follows: STATISTIC, STATISTIC Label, TLIST(A1), Year, C02342V02816, Type of House, UNIT and VALUE.

Type of Variables:

(a) Quantitative variables: TLIST(A1), Year (discrete independent variable), C02342V02816 and VALUE.

(b) Qualitative variables: STATISTIC, STATISTIC Label, Type of House and UNIT.

#### **Data Preparation and EDA**

The most relevant codes used in the Jupiter Notebook for the data preparation and EDA are described below:

* Importing the required libraries using the code ‘import as’ and ‘from import’ so we can gain access to pre-existing codes and materials that can be used to perform arithmetic operations, visualize data, machine learning, etc.

A picture containing text, screenshot, font

Description automatically generated

Figure 1: Required libraries

* Loading the csv file and checking first five (5) rows using the code ‘pd.read\_csv("mycsvfile.csv")’ and ‘.head()’ respectively.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 2: Loading dataset and displaying the first five (5) rows of the dataset

* Using the code: *‘.shape’* to return the number of rows and columns.
* Using the code: ‘*.info()’* to display a shorter summary of the dataset. As show in the figure below, the dataset is composed by 49 rows and 8 columns with no NULL values.

A screenshot of a computer code

Description automatically generated with medium confidence

Figure 3: Structure of the Dataset

* The describe code ‘.describe’ provides purely descriptive information about the dataset. This function is extremely important as will allow you to visualize all the data contained in the dataset.
* The code ‘Print(ppg.columns)’ provides a list of all the columns in the dataset, which can be useful if you need to drop, modify and/or add columns.
* Dropping unnecessary columns: After analysing the data displayed using the ‘.describe’ and Print(ppg.columns)codes, we could see that there were some columns that are not needed for the purpose of the analysis so we decided to drop them from the dataset using the following code: ‘ppg.drop(['STATISTIC', 'TLIST(A1)', 'C02342V02816', 'UNIT',], axis=1, inplace=True’).
* Renaming the remaining columns to normalize the data ‘.rename(columns={'STATISTIC Label': 'Statistic Label', 'VALUE': 'IRL Total'})’.
* Adding new columns using the code: ‘ppg['Country\_1'] = 'Ireland'’.
* Setting the column ‘Year’ as the new index of the dataset using the code: ‘ppg = ppg.set\_index('Year')’.
* Checking the index using the code: ‘print(ppg.index)’.
* Checking for missing values in the data using the code: ‘ppg.isna().sum()’.
* Creating a new cvs file with the index using the code: ‘ppg.to\_csv('/Users/AP139GE/MSc\_CA2/Dataset/NDC.csv', index=True)’.
* Plotting the data with a line graph using go.scatter plot for a more dynamic and descriptive graph as showed in the figure below.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 8: New Dwelling Completions 2011-2022

* 2011: The total number of new dwelling completions in 2011 was 6994.
* 2012: The total number of new dwelling completions in 2012 was 4911, down -29.78% from the 6994 in 2011.
* 2013: The total number of new dwelling completions in 2013 was 4575, down -6,84% from the 4911 in 2012.
* 2014 to 2017: 20,61%, 30.83%, 36.04%, 45.57%
* 2018: 17,872, 25,01%
* 2019: 21,126, 18,21%
* In 2020, despite the COVID restrictions: The total number of new dwelling completions was 20,574, down -2,61% from the 21,126 in 2019.
* The impact of the COVID-19 pandemic and associated restrictions continued to affect completions in 2021, where the total number of new dwelling completions was 20,553, down -0,10% from the 20,574 in 2020.
* The total number of new dwelling completions in 2022 was 29,822, up 45,10% from the 20,553 in 2021. This level of completions shows a huge recovery from the impact of the COVID-19 pandemic and associated restrictions which particularly affected completions in 2021.

From the three types of new houses completed shown in the figure below, the variant apartment was the lowest one until 2020, jumping from 3,909 units in the mentioned year up to 5,126 in 2021, which represents an improvement of 31,13%. Compared with the number of single houses completed in 2021, the variant apartment growth 8.19%, positioning the variant above.

This improvement located this type of constructions above the single house trend in the following year 2022.

On the other hand, the biggest relative growth can be seen for scheme houses, which in 2014 went from 1,795 to 3,294 in 2015, this is 83,51% increase from one year to the other.

which upward the number of single houses trend up 2015 was 3,252, xx% less than the total number of scheme houses.

On the other hand, although in 2011 the number of completions of scheme houses was below single houses, 1,358 and 4,814 respectively, in 2015 the trend of scheme house started to improve significantly, going from 1,795 in 2014 to 3,294 in 2015, this is 83,51% increase from one year to the other.

positioning itself above the other two variants for the following years.

The trend of new Apartments

* Histogram of # of New Dwelling Completions:

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 8: Histogram of New Dwelling Completions 2011-2022

* Descriptive statistics for each type of house:

A picture containing text, font, screenshot, algebra

Description automatically generated

Figure 8: Descriptive statistics for each type of house

* Box plotting descriptive statistics for type of house:

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

Figure 8: Box plot of Descriptive statistics for each type of house

* NDC

A picture containing plot, diagram, line, screenshot

Description automatically generated

Figure 8: Normal Distribution of # of NDC

* KDE

A picture containing plot, diagram, text, screenshot

Description automatically generated

Figure 8: KDE

**Preparing the data**

Basic information about data – EDA

Missing values

Find the Null values

Replace the Null values

Replace the Null values

Filter the Data

box plot

**Analyzing the data**

Bivariate analysis

Multivariate analysis

The most relevant codes used in the notebook for the data preparation and EDA are described below:

* Import the necessary libraries using the ‘**import**’ statement so we can gain access to their functionality and can use them to perform various operations and analyses in our Jupyter notebook. We imported libraries such as Pandas, Seaborn. Numpy, Pyplot, Matplotlib.
* The Pandas package is used for importing the database, using a method called ‘**read\_csv()**’ that allows us to read data from our CSV file and create a DataFrame in Python. Also, we are dropping the index from the dataset using the using the ‘**index\_col=0’**’ method so a new sequential index can be created if needed. In our case, we want the column that contains the years to be reset as our new index.

Table

Description automatically generated

Figure 2: First Five Rows of the Dataset after Dropping the original Index

* Checking the dataset Head and Tail using ‘**df.head()**’ and ‘**df.tail()**’ methods to quickly inspect the structure of the data, and to verify that the data is being read correctly from the csv file. We can appreciate that the features in the dataset are quantitative continuous variables.
* Reset the index, to do so we used the code ‘**df.columns.name = 'Year'**’ to set the name of the column index in a Pandas DataFrame to '**Year'**. And we also check if the change was successful using the '**df.head()**' method.
* Check the data type of each column in the DataFrame using the ‘**dtypes**’ method so we can know the data type of each column. Knowing the data type of our DataFrame is important for performing data analysis tasks such as aggregating data, filtering data, and performing mathematical operations on the data. In our case all our data type is ‘**Object**’, which means that if we want to perform mathematical operations on our dataset, we will not be allowed.

A picture containing table

Description automatically generated

Figure 3: Field Types of the Dataset

* Check the index using ‘**print(df.index)**’ so you can see the labels of the row index for the DataFrame and the data type of the index.
* In our instance, we need to transform objects/strings into integer as we are working with quantitative continuous variables. To do so,
  + first we create a list of the column names with ‘**df.columns.to\_list()**’, and
  + then we iterate over each column to replace commas with nothing using the ‘**replace**’ method with a regex pattern, and then converts the values to integers using the ‘**astype**’ method,
  + Finally, we check our dataset using the ‘**df.head(2)**’ code to display the first two rows of the modified dataset.

Table

Description automatically generated

Figure 4: Updated Dataset

* Check the (‘**df.info()**’) method to display a summary of information about the dataset, so we can quickly inspect the structure of a dataset and identify any missing or null values in the data. From the figure below, we can appreciate that our dataset contains 7 columns and 19 rows, our data type is ‘**float**’ which means that we can perform arithmetic operation on the data so we can use machine learning to implement the regression algorithm.

A picture containing text, receipt

Description automatically generated

Figure 5: Structure of the Dataset

* Use the (‘**df.describe()**’) method to generate descriptive statistics of the dataset.
* Checking whether there are any null values in the dataset or not. As we can see from figure 5, we are not finding any null value.

Table

Description automatically generated

Figure 6: Null Value Detection

#### Data Exploration/Visualization

Exploratory data analysis is carried out on the clean dataset, which results in visualizations and charts. Data Exploratory Analysis is done using building charts and plots. Without performing any Exploratory Analysis, it is impossible to understand the relation with the independent and the dependent features by drawing plots like scatter plot, histogram, bar plots demonstrate their relationship. Data Exploration in this project is supported using Python data visualization libraries such as Seaborn, Pyplot, and a few more. Using these packages for visualizations contributed to interactive plots, making information better to understand. The other purpose that Data Exploration solves is answering the questions that arise by looking at the dataset.

The graphical representation of data can elaborate on the relation of the database. Figure 7 demonstrates a heatmap on the correlation between features. Correlation values represent the interrelation between features with values ranging from -1 to 1. The higher the value, higher the correlation and vice-versa, a positive one interprets a positive relation (increment in one increases the other) and negative one indicates a negative relation (decrement in one decreases the other).

Based on what we can see from Figure 7: Heatmap, most of the correlation values are high, indicating a strong positive correlation between the geographic areas. For example, the correlation between Cork and National is 0.99, which means that when the value for Cork increases, the value for National also tends to increase, and vice versa. The same holds true for the other variables as well, with high correlation values ranging from 0.94 to 1. Overall, the heatmap suggests that the different geographic areas have strong positive correlations with each other, indicating that they are likely to be influenced by similar factors or share common characteristics.

Chart

Description automatically generated

Figure 7: Heatmap of Features Correlation

On figure 8 we have chosen a line graph chart in Python using Matplotlib to plot the average house price per year in Ireland from 1997 to 2015. This is because line graphs are especially useful to compare variables that increase over time so we can compare changes over the same period of time for more than one variable.

Unsurprisingly, Dublin has the most expensive residential properties greater than the six other variables presented in the graph, the gap between Dublin and the rest of the variables is wide. The graph shows that the house prices in Ireland experienced a significant increase during the late 1990s and early 2000s, with a peak in 2007. After the peak in 2007, the prices decreased, and the market experienced a downturn, particularly in Dublin, Cork, and Galway. However, the prices started to recover in 2013 and continued to rise until 2015.

The national average price of houses in Ireland in 1997 was around 102,000 Euros, while the average price in 2015 was around 282,000 Euros. Dublin, the capital city, had the highest average house prices throughout the period, with an average price of around 123,000 Euros in 1997, rising to around 381,000 Euros in 2015. Cork, Galway, Limerick, and Waterford also experienced significant increases in house prices, but they were lower than those in Dublin. In conclusion, the line graph shows that the house prices in Ireland experienced significant fluctuations over the period, with a sharp increase in the early 2000s, a peak in 2007, a decline afterward, and a recovery in recent years.

Chart, line chart

Description automatically generated

Figure 8: Average House Price in Ireland (1997-2015)

Between 1997, the first year for which we have median sales price data, and 2001, home sales prices increased 78% nationally. All the country’s major counties experienced strong house price rises during the period from 1998 to 2001, being Dublin the city that experienced the biggest increase during the mentioned period. However, from the table below, you can also appreciate a significant drop on the percentage of increase starting on 2007, this is due the presence of a credit bubble in the early to mid-2000’s. The collapse of the credit bubble, owing in part to the emergence of the global financial crisis of 2007/08. It wasn’t until 2014 that the real state market started to recover.

Table

Description automatically generated

Figure 9: Percentage of difference between one year and the next

Looking at Figure 10, the boxplot that plots the descriptive statistics of the dataset, we can appreciate that Dublin has the highest median and highest overall mean house prices, followed by Cork and Galway. Waterford has the lowest median and lowest overall mean house prices, followed by Limerick and Other Areas. There are no significant outliers in the data, indicating that the mean house prices for each city are relatively consistent over the time period covered in the dataset. Overall, the box plot shows a clear and concise summary of the distribution of mean house prices by city, which allow us to identify which cities have higher or lower house prices, as well as any significant outliers in the data.

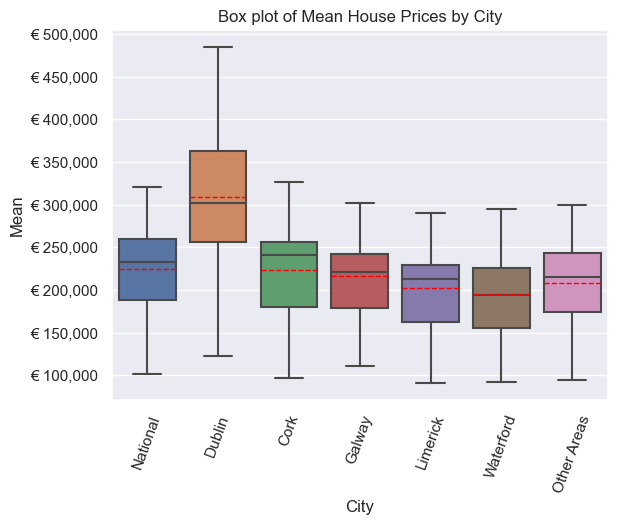


Figure 10: Box plot of Mean House Prices by City

### Statistics: Distribution

One distribution that could be used to explain/identify some information about the dataset is the Normal distribution. A normal distribution of house prices in Dublin based on the given mean and standard deviation. The mean value of 309,492 represents the center of the distribution, while the standard deviation of 97,607 indicates the spread of the distribution. These values are used to define the parameters of the normal distribution.

The resulting plot below shows a bell-shaped curve, which is a characteristic shape of a normal distribution. The x-axis represents the price range in euros and the y-axis represents the probability density. The highest point of the curve represents the most common price value in the dataset. The curve tapers off to the left and right, indicating the decreasing probability of getting a price that is further away from the mean value.

Chart, line chart

Description automatically generated

Figure 11: Box plot of Mean House Prices by City

Figure 12 shows the Density Estimate (KDE) plot of the distribution of prices for houses in Dublin using the Seaborn library. In this case, the x-axis represents the range of house prices in Dublin, and the y-axis represents the density of houses in that price range. The KDE plot shows that the distribution of house prices in Dublin is roughly symmetric and centered around the average price of approximately €309,492.

Chart

Description automatically generated

Figure 12: KDE Plot of Dublin

The figure below shows the probability that a new house in Dublin will have a sale price of less than or equal to €400000 is 0.809 or 80.92% approximately.

Text

Description automatically generated

Figure 13: Calculates the probability that a new house in Dublin with a price of €400000 or less will be sold

To calculate the probability of selling a new house in Dublin for more than €400,000, we calculate the z-score of 400000 based on the mean and standard deviation of house prices in Dublin using the formula:

z-score = (400000 - mean) / standard deviation

We use a standard normal distribution table or calculator to find the probability of a z-score greater than or equal to the z-score you calculated.

The probability of selling a new house in Dublin for more than €400000 is 0.1769

Graphical user interface, text, application, email

Description automatically generated

Figure 14: Calculates the probability that a new house in Dublin being sold for more than €400000

### Model Building and Results

#### Model Building

There are three types of machine learning algorithms we were asked to choose from to use for our project: supervised, semi-supervised and unsupervised.

1. Supervised machine learning is defined as machine learning done with labelled data. Labelled data could for instance be a dataset containing attributes describing a set of cakes and a label for each cake indicating whether the cake tastes good or not.
2. Unsupervised machine learning is learning without a label set. The learner does not know specifically what prediction is “correct” and cannot correct itself according to labels as supervised learning can.
3. Semi-supervised learning is similar to supervised learning, but instead uses both labelled and unlabelled data. Labelled data is essentially information that has meaningful tags so that the algorithm can understand the data, whilst unlabelled data lacks that information. By using this combination, machine learning algorithms can learn to label unlabelled data.

The type of machine learning we will be operating within is supervised machine learning. In the case of real estate price prediction, the label is tied to price of the houses and the rest of the data can be attributes describing the year.

When creating a machine learning model, the input given to the learner can be divided into different types of data (Shalev-Shwartz and Ben-David, 2014). The learner is given a training set of data, which contains a domain set of data and a label set of data. The domain set is the data describing attributes believed to be predictive of the label set. The label entries are tied to specific entries in the domain set.

RESEARCH QUESTIONS

1. Create a model to predict property pricing.

2. What are the features that influence the property pricing?

METHODOLOGY

The CRISP-DM process is used to achieve the project's goal. CRISP-DM stands for Cross

Industry Standard Process for Data Mining. The CRISP-DM methodology is used for

structured data mining project planning. CRISP-DM is divided into six stages.

RESEARCH QUESTIONS

1. Create a model to predict property pricing.

2. What are the features that influence the property pricing?

METHODOLOGY

The CRISP-DM process is used to achieve the project's goal. CRISP-DM stands for Cross

Industry Standard Process for Data Mining. The CRISP-DM methodology is used for

structured data mining project planning. CRISP-DM is divided into six stages.

RESEARCH QUESTIONS

1. Create a model to predict property pricing.

2. What are the features that influence the property pricing?

METHODOLOGY

The CRISP-DM process is used to achieve the project's goal. CRISP-DM stands for Cross

Industry Standard Process for Data Mining. The CRISP-DM methodology is used for

structured data mining project planning. CRISP-DM is divided into six stages.

RESEARCH QUESTIONS

1. Create a model to predict property pricing.

2. What are the features that influence the property pricing?

METHODOLOGY

The CRISP-DM process is used to achieve the project's goal. CRISP-DM stands for Cross

Industry Standard Process for Data Mining. The CRISP-DM methodology is used for

structured data mining project planning. CRISP-DM is divided into six stages.

RESEARCH QUESTIONS

1. Create a model to predict property pricing.

2. What are the features that influence the property pricing?

METHODOLOGY

The CRISP-DM process is used to achieve the project's goal. CRISP-DM stands for Cross

Industry Standard Process for Data Mining. The CRISP-DM methodology is used for

structured data mining project planning. CRISP-DM is divided into six stages.

The output of the learner is a prediction rule that can be used to label new data. The success of a trained model can be measured as the error of predictions made to new data. The less the error, the better the model.

Under the umbrella of supervised learning fall: Classification, Regression and Forecasting.

* Classification: In classification tasks, the machine learning program must draw a conclusion from observed values and determine to what category new observations belong. For example, when filtering emails as ‘spam’ or ‘not spam’, the program must look at existing observational data and filter the emails accordingly.
* Regression: In regression tasks, the machine learning program must estimate – and understand – the relationships among variables. Regression analysis focuses on one dependent variable and a series of other changing variables – making it particularly useful for prediction and forecasting.
* Forecasting: Forecasting is the process of making predictions about the future based on the past and present data and is commonly used to analyse trends. A method of forecasting and now-casting of prices in each market is by extrapolation using a price index. The price index works by using the average market trend to extrapolate a previous sales price to current time. The index can be described as:

Based on the above, the approach that we will take is Regression and Forecast since they are the ones that best fit the task of predicting the price of a house with an error as small as possible.

The mean squared error (MSE) is a measure of the average squared difference between the predicted and actual values. In this case, the MSE value of 11204648207.808912 suggests that the model's predictions are not very accurate, as there is a large difference between the predicted and actual values.

The predicted prices for the years 2016 to 2020 are shown as an array of five values, each representing the predicted price for the corresponding year. The predicted prices are based on the trained model, which uses the relationship between the year and house prices in Dublin to make these predictions.

For example, the predicted price for 2016 is 392399.80 euros, which means that the model predicts that the average price of a house in Dublin in 2016 would be around 392399.80 euros. Similarly, the predicted prices for 2017, 2018, 2019, and 2020 are 401788.56, 411177.32, 420566.08, and 429954.85 euros, respectively.

Graphical user interface, text, application, email

Description automatically generated

Figure 15: Machine Learning

Regarding hyperparameter tuning, for the Linear Regression model, there is only one hyperparameter to tune, which is the regularization parameter alpha. We can use GridSearchCV or RandomizedSearchCV to find the optimal value for alpha. Using GridSearchCV, GridSearchCV searches for the best value of alpha from a list of possible values. The cv parameter specifies the number of cross-validation folds to use. After fitting the GridSearchCV object, you can access the best hyperparameters using the best\_params\_ attribute. You can then use the best hyperparameters to train the final model on the full training set and evaluate its performance on the testing set.

Graphical user interface, text, application

Description automatically generated

Figure 16: GridSearchCV

The output "Best hyperparameters: {'alpha': 100}" suggests that the best value for the hyperparameter 'alpha' was found to be 100 using the hyperparameter tuning method (either GridSearchCV or RandomizedSearchCV).

The hyperparameter 'alpha' is used in the Ridge regression algorithm, which is a variant of linear regression that adds a regularization term to the loss function to prevent overfitting. The 'alpha' parameter controls the strength of the regularization, with higher values of 'alpha' resulting in stronger regularization and lower values resulting in weaker regularization.

The fact that the best value of 'alpha' found was 100 indicates that a relatively strong regularization was needed to achieve the best performance on the data. This could suggest that the data was prone to overfitting and that stronger regularization was needed to prevent this.

Overall, the best hyperparameter value of 'alpha' found using hyperparameter tuning suggests that a Ridge regression model with strong regularization would perform well on the data. However, it's important to note that this is just one possible hyperparameter combination and other combinations could also result in good performance.

### Evaluation

In this section, we will present our results and present the insights gained from our experiment with machine learning and forecasting where we have used real estate sales data. As a part of the CRISP-DM process, this represents the “Evaluation” element.

This section provides an overview of the two models used by evaluating them on the basis of the metrics.

* Linear Regression Model: The first tested model was built using Linear Regression. Despite of being aware of the facts that Linear regression models cannot deal with non-linear relationships, this model was implemented to see how much better it performs than the state of the art work models on similar dataset.

|  |  |  |
| --- | --- | --- |
| Model | Actual Value | Predicted Value |
| Regression |  |  |
| Forecasting |  |  |

Table 1: Actual vs Predicted Prices

### Deployment

The deployment process of the CRISP-DM model was not considered as a part of the scope for this project. That being said, some comments will be made on the process since we have used the CRISP-DM process extensively. The CRISP-DM model was used as a framework for the experimentation.

After creating a model with satisfying prediction accuracy, deployment is an option. As a part of the CRISP-DM, deployment is started when satisfying value creation has been validated through the previous process steps of “Business Understanding”, “Data understanding”, “Data preparation”, “Modelling” and “Evaluation”. In order to keep the model used in deployment up to date, it needs to be retrained continuously. Compared to the stock market or the cryptocurrency market, the real estate market is not sensitive on a day-to-day basis. This means that a machine learning model deploying in a production environment does not constantly need to retrain with new data in order to stay up to date. This has the effect on the deployment, that retraining the model with new data only has to happen on daily or weekly basis and still produce predictions with a consistent accuracy. This would of course have to be tested and validated, since it is only an assumption. In a production environment, it would be possible to use web services to connect to the trained model in Azure Machine Learning and feed predictions to the application using the trained model.

## Conclusion

The widespread acknowledgement that actual housing completions in the Irish housing

## Acknowledgements

I would like to thank my family for being support during this stressful time.

## References

The department of [housing](https://moodle.cct.ie/mod/resource/view.php?id=129150) local government and heritage in the domain of [HOUSING](https://moodle.cct.ie/mod/resource/view.php?id=129150), This data is available at: <https://data.gov.ie/dataset?q=houses&organization=department-of-housing-local-government-and-heritage&sort=score+desc%2C+metadata_created+desc&theme=Housing>