Specialist Certificate in Data Analytics Essentials - Final Course Project

https://github.com/drmarbles/UCDPA_atracey

Abstract

The Irish housing market has been a major topic of conversation in this country for decades - from the boom of the Celtic Tiger years, to the following bust, and subsequent recovery. Prices for the same property can fluctuate wildly over time but what factors go into determining the price of a property? Can the price be accurately predicted? That is the aim of this project, along with outlining interesting insights from a dataset of Irish property data, looking at how the dataset splits by various metrics.

Introduction

From reading course material and also researching potential datasets online, I had seen that house price prediction was mentioned as a good project, as there is a clear outcome (price) that can be predicted from a variety of suitable variables. I came across the Daft House Price dataset on Kaggle and it seemed ideal as it covered the topic and also had a local focus. There was also a personal connection as, having recently bought a house, I have spent a lot of time on Daft and so feel I know it, and then Irish housing market, well.

Dataset

The dataset was posted on Kaggle (https://www.kaggle.com/datasets/eavannan/daftie-house-price-data). It was extracted from the Daft website and covers the period Jan 1st 2020 - Jan 30th 2022. There are 3967 properties in the dataset. The data was made available as a CSV file containing 22 different fields.

Field	**Description**
id	Unique property identifier
title	Property address
featuredLevel	The ad's level on the website - can be standard, featured or premium
publishDate	Date of publication of the listing
price	Property's listed price
numBedrooms	Number of bedrooms
numBathrooms	Number of bathrooms
propertyType	What kind of structure is the property? Possible values - 'End of Terrace', 'Semi-D', 'Terrace', 'Detached', 'Apartment', 'Bungalow', 'Townhouse', 'Duplex', 'Site', 'Studio' ,'House'

propertySize	Size of the property in square metres
category	Either ‘Buy’ or ‘New Homes’ - simply describes which section of the website the ad appears under.
AMV_price	Advised minimum value
sellerID	Unique identifier for the seller
seller_name	Seller’s name
seller_branch	If an agency, agency branch
sellerType	The type of seller. Possible values - ['BRANDEDAGENT', 'UNBRANDEDAGENT', 'PRIVATEUSER']
m_totalImages	Total number of images posted on the website
m_hasVideo	Whether the property has a video on the website or not
m_hasVirtualTour	Whether the property has a virtual tour on the website or not
m_hasBrochure	Whether the property has a brochure on the website or not
ber_rating	BER rating for the property
longitude	Property location geocoordinate
latitude	Property location geocoordinate

The dataset appeared to be a well structured set of data, with clear, logical columns and so should prove to be a good, interesting dataset to work with. It should be noted that the price field listed here is not the price that a property was sold for, but the price that the property was listed for on Daft.

Implementation Process

Whilst the dataset could be manually downloaded from Kaggle through a browser, I decided to download it using Kaggle’s API, as this would be good practice for similar projects in the future (and also because it covered one of the project requirements).

I loaded it into a dataframe and did some inspection of the data. Many of the fields are particular to the listing of the property on Daft’s website and so do not contain information about the property itself. The most pertinent for this project were - title, price, numBedrooms, numBathrooms, propertyType, propertySize and ber_rating.

I know from my personal experience of the Irish property market that location is a key determinant of price, and so wanted a level of granularity for location that could be used to group together properties. One option would have been to use longitude and latitude to establish precise location using a lookup module such as Geopy, however I decided that I could easily get the county from the information in the title field. The addresses all seemed to have similar formats, with the county information being all the text after the last comma in the field. I wrote some regex to extract this into a column ‘county’ in a new dataframe.

Upon viewing the list of extracted counties, there were some entries which did not contain the correct information. Though small in number (so I could have just dropped them from the dataset) it was easy enough to replace each value with the correct county information, albeit in a manual and in no way programmatic way.

Properties in Dublin were listed at the postcode level and I decided to keep this level of granularity here as a) Dublin is so much more populous than the other counties and so I wouldn't be diluting the data too much and b) again, from personal experience I know that Dublin postcodes differ hugely in house prices.

I merged these two datasets together so the original dataset now had county level data.

I knew that I would be creating quite a few graphs to visualise the data so decided to create a function to do that. The function created a bar graph plotting the number of properties in each category that was passed through to it, with the output being the required graph. I did a similar process with graphs that use average house price as the metric to be plotted.

Before moving onto the modelling part of the project I examined the dataset for missing values and found that the propertySize field contained 355 null values out of the total of 3967. This meant 9% of the values were missing. I did not want to just remove these rows from the dataset as 9% is a substantial number and I wanted to preserve the data from the other fields, so I replaced the entry for each of these rows with the median property size of the rest of the dataset.

For the modelling I tried two different models. The first was a Random Forest, and the second was XGBoost. The reason for choosing these is that I was attempting to do a regression from multiple inputs and my reading of modelling techniques suggested these would be two good options to test.

Firstly I ran the random forest regression, using numBedrooms, numBathrooms, propertyType, propertySize, ber_rating and county as my input variables. I felt these were the ones in the dataset that would have an impact on price - many of the others are concerned with the listing of the property on Daft. County, propertyType and ber_rating had to first be encoded as numbers in order for the model to work. The result of the run was a root mean square error of 166,747.

I next tried an XGBoost model, with the same input variables. This returned a root mean square error of 159,252, which was slightly better than the random forest, though considering the mean price in the dataset is 341,000 this does not seem like a brilliant prediction.

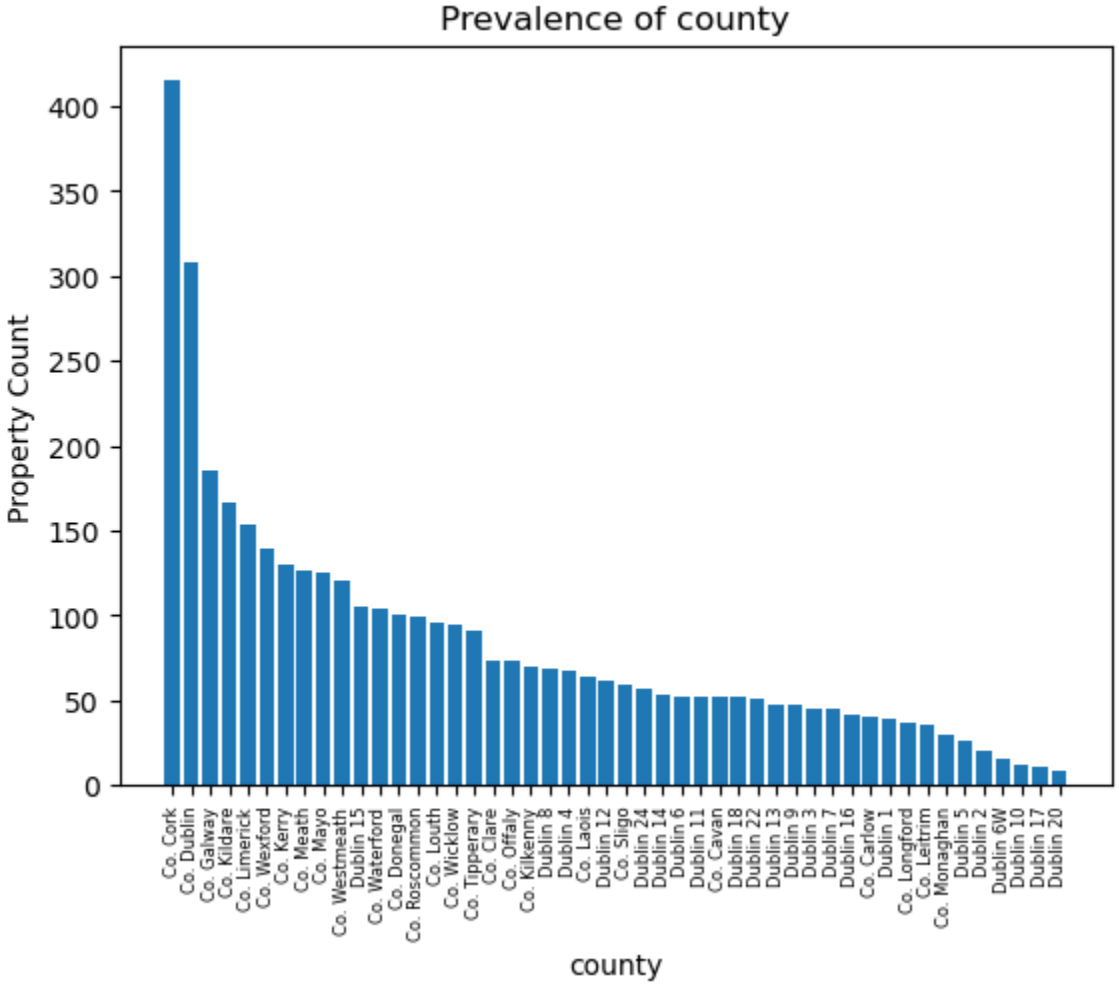
As the XGBoost model had performed slightly better I used that one to carry out some hyperparameter tuning. From my knowledge of the model, it seems that some of the most common parameters to adjust are learning_rate, max_depth and n_estimators, so I put in some additional values in addition to the defaults. I used a random search rather than a grid search purely for computational reasons - grid search was taking far too long to run.

Running the model now produced a root mean square error of 157,785, so slightly better. The best parameters found were n_estimators: 300 (default is 100), learning_rate: 0.1 (default is 0.3) and max_depth: 5 (default is 6).

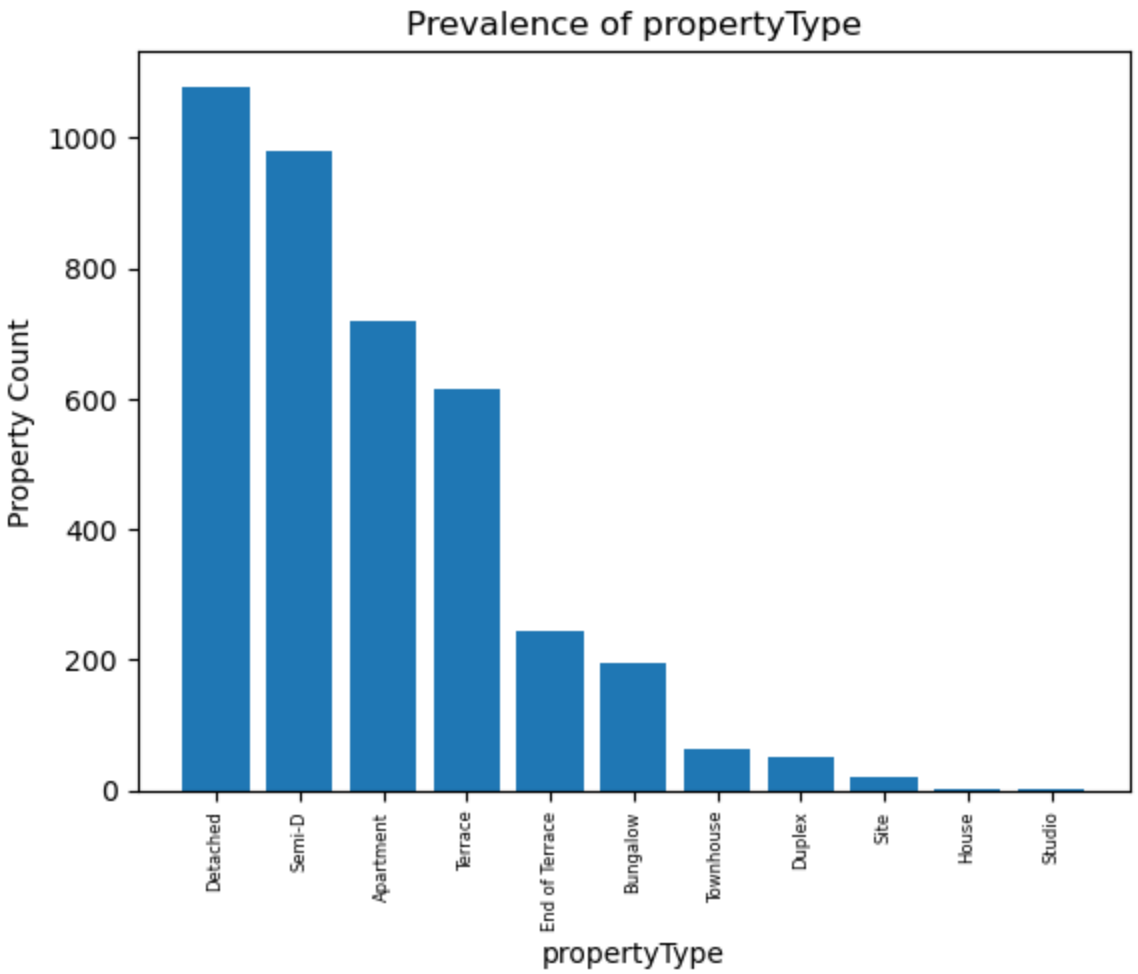
I would have liked to have seen a better performance from the model as I think an error in the region of 50% of the target variable is not that great - it would be interesting to run a similar project with a wider range of data, perhaps in a larger country.

Results and Insights

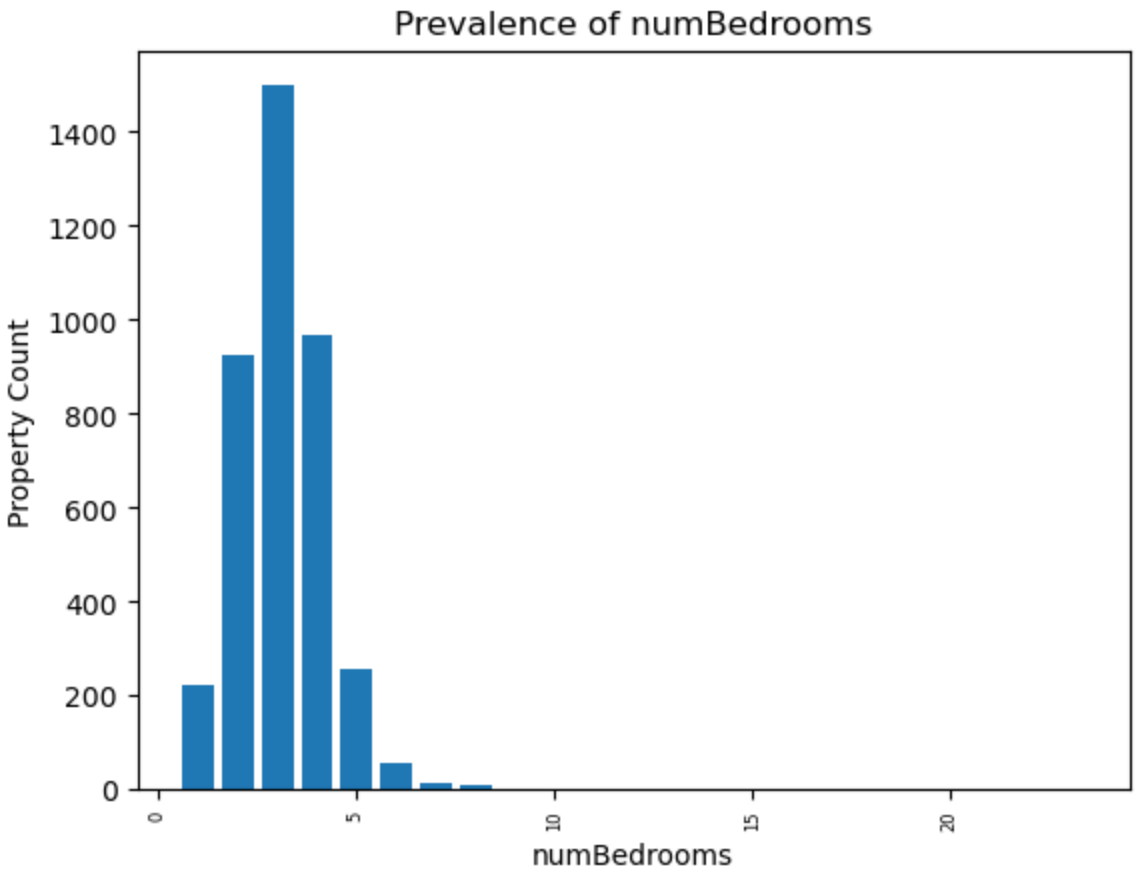
- Cork and County Dublin were the areas which had the most properties listed on the website.



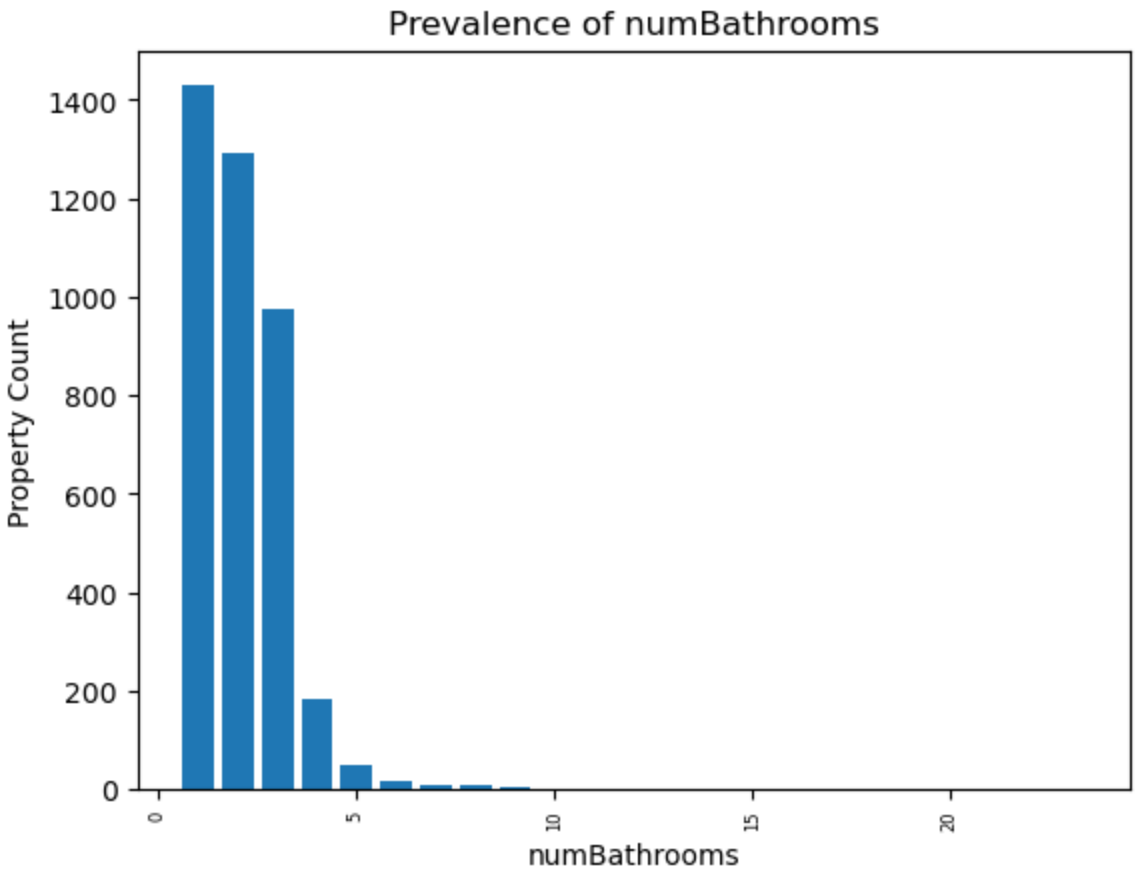
- Detached was the property type with the highest number of properties in the dataset, followed by semi-detached. There were very few properties listed as either House (2) or Studio (1).



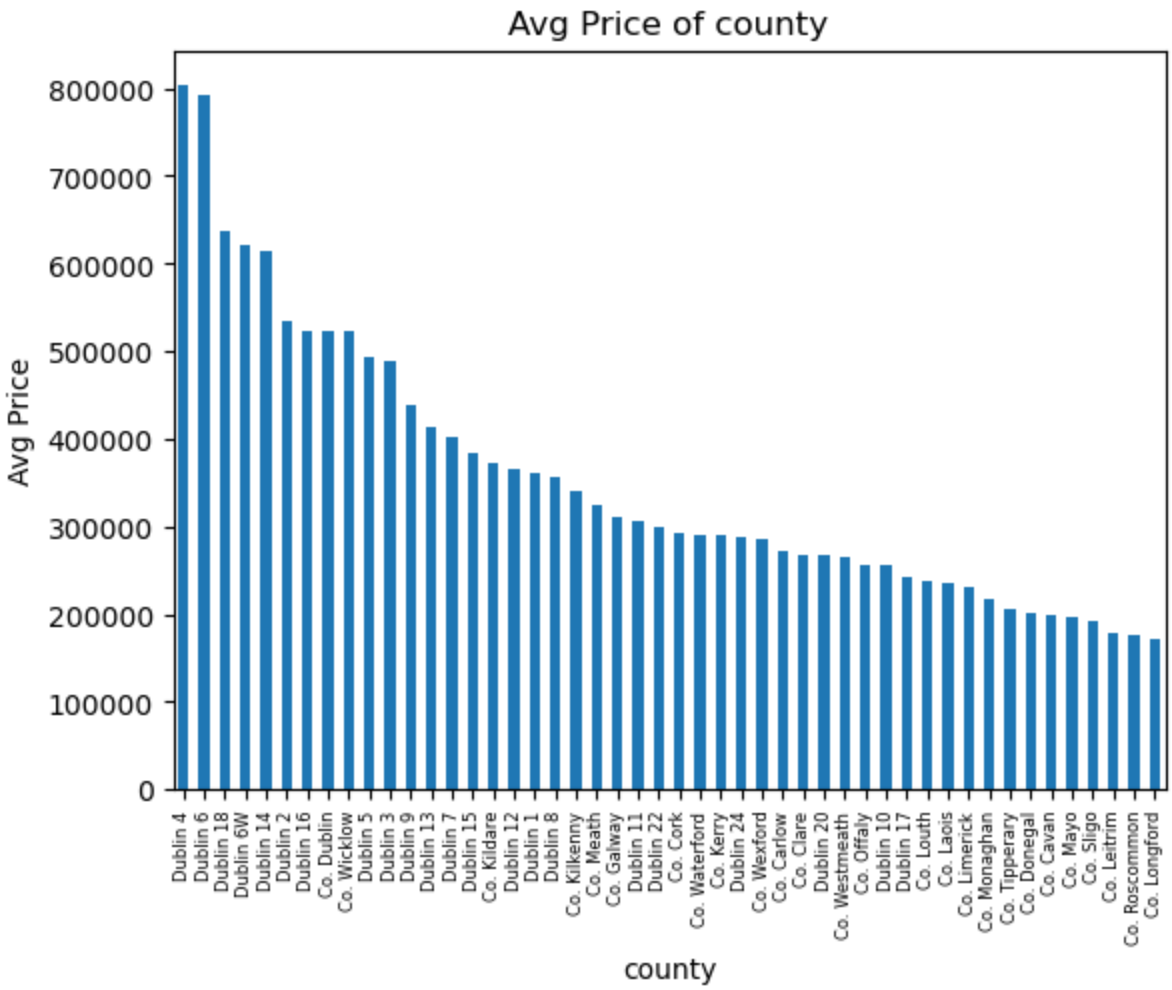
- The most common size of property was 3 bedrooms, with over 1400 properties listed. Four bedroom and two bedroom properties were next with around 950 each.



- The most prevalent number of bathrooms was 1 which, when combined with the information from the previous graph suggests many of these properties are fairly old, as newer properties have a much more correlated number of bathrooms to bedrooms as people today don't want to wait for the bathroom if there's someone in there (like I had to when I was a kid growing up in our 3-bed, 1-bathroom house 🙂)



- The most expensive areas in the dataset by far were Dublin 4 and Dublin 6. This certainly tallies up with my experience of the Dublin housing market. The most expensive area outside of Dublin was Wicklow. The cheapest county in Ireland is Longford.



References

Daft.ie house price data set, uploaded to Kaggle - https://www.kaggle.com/datasets/eavannan/daftie-house-price-data