

Property Price Prediction

Alan Cahill

C00174533

Lecturer:

Dr. Lei Shi

Project Report

**Table of Contents**

**Abstract3**

**Project Description4**

Crawler**4**

Pre-Processing**5**

Classifier**6**

Q-Learning**7**

**Description of Learning8**

Personal**8**

Technical**10**

**Review of Project12**

What Went Well**12**

What Didn’t Go Well**12**

What Would You Change**13**

What Future Work Can Be Done**13**

**Abstract**

The purpose of this document is to explain the final iteration of the project or the project in its completed form at the time of the final demonstration. The document shall explain the project from a technical perspective while also explaining the problems faced and the solutions to these problems. Along with this will be the explanation of the learning that was achieved throughout the project, both technical and personal. Finally, the document will review the project and discuss ideas that could be implemented in further development of the project.

**Project Description**

The project itself is comprised of the following aspects:

* A crawler
* Pre-processing
* A classifier
* A Q-learning implementation
* Stats and Graph jupyter notebook

These aspects mentioned are all necessary to achieve the project goal of property price prediction using machine learning. The final aspect mentioned, Stats and graph jupyter notebook, is the only non-essential aspect. To attempt property price prediction without utilising statistics and graphs would be almost cheating. All the decisions made throughout this project were decided by assessing statistics or graphs. For this reason, the stats and graphs jupyter notebook is actually a fundamental aspect of this project and without its use could have affected the accuracy and the direction of the project.

**Crawler**

To achieve machine learning for property price prediction a dataset is the backbone of the process. Without a dataset then the project could not even get off the ground. Unfortunately, there was no dataset provided at the project’s initiation. It was expected that the data set would be gleaned by the project engineer. The first place a datasets availability was researched was the property price register. This was mildly successful as the data here was easily downloadable and provided for every property sold in Ireland in the last 10 years and more. This data set only provided an address, a type, a price and a date. This data was not nearly sufficient enough for completion of the project. This lead the project in the direction of online property advertisement websites. The one chosen was Daft.ie. There was no technical reason for choosing Daft as every website researched provided almost the very same data on property registered with them. Daft however provided more properties, or in other words, a bigger dataset. This data was not downloadable. Neither was the data from the other websites. To retrieve the data, it would have to be gleaned through reading the website and entering the data into a database. This would be foolish to do manually. Instead, there exists a method for achieving exactly that in an automated program. This way is called a crawler. A crawler is a software device that has been created to read the website and retrieve the necessary information and placing it in a dataset. The reading and retrieving of the data is carried out by software called a scraper. This software device parses the html source of the webpage and searches out and saves the specified data to a dataset. The scraper is only capable of reading the html of one web page at a time. For this reason, the crawler is necessary as the crawler will traverse the webpage for the scraper aspect of itself. On each page there are links. These links lead to every single area of the website. This project however only required access to certain areas. This area is the properties listed for sale in Ireland. This exact area was chosen as the starting point of the crawler. On the page in question there exits all kinds of links. These links bring the website to the home page, rent page and help page to name a few. These however are not the areas the crawler needs to get to. There are 20 pictures on this page and each of these pictures bring the user to the individual Daft page for the property it represents using a link embedded into the picture. At the bottom of this page there is a ‘next’ button. This button brings the user to the next 20 properties. The crawler uses a second scraper implementation to scrape the 20 links and appends them to a python list in the backend. Once this is done the crawler will retrieve the link behind ‘next’ button and will traverse to this webpage to begin the process of scraping links and appending them to the list all over again. This continues until all the properties links have been retrieved from daft. Once all the links have been appended to the python list then the crawler will traverse this list and will use each appended link to visit the corresponding daft.ie webpage. Once on each individual link on daft.ie the crawler will utilise a different scraper implementation to scrape/retrieve the data that is needed from each webpage. This data is incrementally added to a dictionary and then once all properties have been retrieved the data is transferred to a database. This is the final act of the crawler.

**Pre-processing**

The crawler achieves its goal of retrieving the data for the betterment of the project. The data is exactly what was needed for continuing the development of the project overall. However, the data was not exactly perfect. The data retrieved was as seen in Fig.1.DATABASE PIC HERE. The data included a price, an address, a bedroom count, a bathroom count, a property type, views, history of prices, the dates the prices were changed a living area in squared meters and details. As seen in Fig.1 the data in the database is not ready for processing, with the ‘Price On Application’ inside the price column at position 5 being a prime example. For the purpose of fully explaining the final iteration of the pre-processing program in this project the document will not talk through the process of getting from the current data as in fig to the final data that was passed into the classification algorithm as shown in Fig.2. Instead the document will explain what the pre-processing program achieves. This sounds as if is the exact same thing but throughout the exploration of the data there was numerous upon numerous iterations of the pre-processing program. Each iteration gleaned more information that was assessed and tested. If the information was deemed important or worthwhile it was kept along with any software aspects that were utilised to achieve the information. However, if it was not useful then it was scraped along with the software aspects utilised and will not appear in this document. Only the remaining functionality which was deemed useful will be discussed. First and foremost, the data arrives from the crawler with usually a figure in the region of 30,000 properties saved in a database. The pre-processing program takes all of these properties and adds them to a total houses table. This table contains every single property that was scraped off of daft and includes duplicates as some properties will be still on daft.ie from one crawler run to the next. A third table is then created which remove all duplicates in the total houses table. This third table is processed and converted to a csv file. The csv file makes the data more portable and much faster to read into a python program. With this csv file the data is again loaded to a python program and is processed to ensure the data is capable of being used in machine learning properly. This processing removes all null values, removes all ‘price on application instances’, formats the data properly, removes the views, allprices and pricedates columns, extracts the necessary data from the details column and then finally removes the detail column. Everything that this program achieves is achieved using custom methods built in python without the help of outside libraries or plug-ins. Once all of this has been achieved then the data is saved to a new csv file. This new csv file is then read into a new python program. This program splits up the data into lists of properties based on which county they reside in. Each of these lists is saved to new csv files named after the county of their residence. The next program only required the Dublin and Cork csv files. Each file is respectively read into a python program and the same process is undertaken using each dataset. The data is split up again with either Dublin 15 or cork city centre properties allowed to progress to the next step. This step entails using the address of the property. The address of the property is passed to the google maps API through a python library. The returned data is parsed to find the geo location or the latitude and longitude of the data. The latitude and longitude are added to a column for each property in the data set that can be located by google. This latitude and longitude are then cross referenced with the latitude and longitudes found for the property register dataset. If there is a match the price of sale and date of sale is added to the daft property. This data is now ready to be passed to the classifier algorithm.

**Classifier**

The classifiers job is to identify sub areas that contain properties located in close vicinity to each other and have similar prices. These sub areas are created with only the same type of property in each one. Each property of the type detached house is put in a separate data set and the algorithm creates sub areas that contains properties located in close vicinity to each other and have similar prices based off this dataset. This means for each property type there exist a different makeup of sub areas. Each property is assigned a label which represents the sub area that it resides within.

Algorithm

The classifier is itself a machine learning algorithm. It utilises an algorithmic approach to classify properties. The concept is based off un labelled clustering. Whereby the data is unlabelled and each time the algorithm is utilised the algorithm will unable any previous labels and then proceed to use all of the data available to classify all the properties. What this means is that the labels may change each iteration or when new data has been added, much like the centre points in a k-means clustering algorithm change when new data is added.

Linear algebra

The algorithm utilises linear algebra to achieve its goals. Namely the algorithm utilises the equation of a line in slope-intercept form. This allows the algorithm to section off the map into subsections which makes it easier to ascertain if a property falls within the expected constraints. The algorithm utilises the functionality that allows the finding of which side of a line a point is on using the slope intercept equation of the line in question. Once each property is sorted into a subsection the algorithm proceeds to the next section.

Gate

Each subsection is sorted in order of closest to the centre point. The subsection is then traversed, and each property’s price is compared to that of a gate. The gate consists of a high and a low point. If the property falls within this gate then it is considered to be acceptable. The traversing of each section continues until two properties in a row do not fall between the gate. At this point the border for this sub section is assigned as the final point that falls within the gate or as the centre point if not point falls within the gate. The algorithm repeats this process until all the subsections have been assigned a border point.

Area check

Once all subsections have been assigned border points then an area has been created. This area must be checked for viability. An area can only be considered viable if it contains a specific number of properties which is calculated relative to the size of the data set entered to the algorithm. If an area is accepted then the algorithm continues forward. Once there is a coverage of minimum 90% achieved then the algorithm has completed its classifying. Before the algorithm will accept its end result a test to see the amount of areas within the larger area has been assigned. There must be a minimum of three areas otherwise the result will be rejected and the gate will be set to be larger and the algorithm will begin again. If the algorithm exceeds 10 areas the algorithm will actually automatically reset with a larger gate. If the result is acceptable after the checks then the resulting data is saved to a csv file with the addition of the label on each of the properties.

K-Means Clustering

As a 90% coverage was assigned as an acceptable level of coverage there exists a further 10% that do not possess a label. This 10% must be assigned a label. To achieve this a clustering implementation mirroring k-means clustering was utilised. To summarize whichever centre point one of the 10% is closest to is the area that the property will be assigned.

**Q-Learning**

Environment

The environment within the q-learning algorithm was essential to the workings of the algorithm. The environment consisted of a two dimensional array of dictionaries that was initialised using a list of lists. The dictionaries themselves were initialised in such a way that they created linked list. This linked list was four dimensional doubly linked.

Data Flow

The data flow of the q-learning needed management. There are a total exceeding five thousand combinations of the criteria being used as identifiers and variable for predicting house price value. These are bedrooms, Bathrooms and label. House type is also considered. The properties are saved to house type files. Thus it is easy to account for type in the data flow. With five thousand multiplied by the number of house types the count exceeds 30,000. Environments are created for each combination of the four variables. Any environment that has a house to match have the house appended to a list. Any combination without a house appended to their list are removed, thus reducing the work for the algorithm.

Prediction Mechanism.

Each q-learning algorithm is run through until there are no more possible updates to the environment. To retrieve the prediction the environments are traversed again without updates until an exit found. This exit point is the prediction.

Stats And Graphs

The statistics and graphs are implemented in a jupyter notebook. They follow a storyline, the storyline being that of the projects progression. Starting in All Of Ireland and whittling down to Dublin 15 and then Dublin 15 sectioned by the classifier the notebook demonstrates a statistical journey. The journey demonstrates how the project progressed through its pre-processing and how the final piece of pre-processing(The classifier) performed.

**Description of learning**

Throughout the project there was huge growth and learning achieved by this author. This growth and learning was both technical and personal. Despite the huge array of new technical abilities learned the personal growth was just as substantial in size and importance and may in the future prove to be the more important and useful of the two. A lot of both the technical and personal growth will definitely prove valuable for the author when pursuing a career but also on a personal and spiritual level. Some of the personal growth achieved in this projects lifetime is essential for success and even personal stability in the future.

**Personal**

**Work ethic**

This project as an entire entity was very vast and covered numerous topics. Some of the topics in the project overlapped or complimented each other. This extracted from the work load somewhat. There were other topics that were maybe similar but fundamentally different. One such example is predicting the price a property will be sold at and predicting the price of a property for a period of time into the future. However the two most difficult and vast parts of the project were in the same group (Maths) but were two completely separate branches of maths. The dis-similarity of these branches provided a difficult challenge to undertake simultaneously, and with the addition of the depth required to fulfil the needs of the project, it was the work ethic of the author that really got the project over the line. The authors work ethic was demonstrated fully between the first and second presentations where a lot of ground was covered in learning with more depth, how to develop an algorithm. This was learned in a practical manner, almost trial and error while writing code. Once this was learned the author was able to complete the bulk of the classifier between these two presentations.

The author also learned about a new aspect of work ethic. When working on the data science research the project required the author continually researched and learned information that would prove irrelevant to the project. The research involved required many hours of studying and amounted to no immediate benefit for the author. The effort required to continue to learn and study at the same pace required a deeper work ethic than any work ethic required to achieve a goal that is definitively guaranteed to provide a reward of substance to the worker. Simply put, it is difficult to motivate oneself to continue the work with no reward.

**Researching**

Researching may sound like it could reside within the technical area, but due to the usability of the topic in the authors personal life it was left here. When researching the various area for the project the author was beginning at a beginner or novice level for most except a few of the areas. The author a beginner at researching also just dove head first into the researching of topics. The author began reading any book that could be found on data science. When this did not work the author began watching videos on YouTube. This yielded some success and by looking at some beginner videos the author was able to achieve an starting insight into data science. Long story, short, the author utilised these beginner videos by referencing back to them whilst watching more advanced videos. Once the authors knowledge increased the author began watching expert videos and utilising the intermediate videos as reference points. With enough knowledge accrued to begin with the author also returned to find a book. This book was BOOK HERE and was utilised in synchronicity with YouTube videos. Ultimately a book that suited the author turned out to be the best source of information.

The research process of the author can be applied to personal aspects of the authors life. This research process will assist the author in making some personal improvements. The author can utilise the process to also make some recreational endeavours easier to learn and master.

**An increased interest in Information Technology**

An increase in the interest of the author in Information Technology took a huge role in the authors life at several stages of throughout the year. Of course, this topic is not as impressive as the other topics within the section. The increased interest did however change some aspects of the authors personal life. Without exhausting the subject too much one area the author grew a strong interest in is computer architecture. The author’s interest was really peaked when learning about the hardware makeup of a computer and laptops. This was not part of the project but just an endeavour the author choose to learn about on personal time.

**Ambition and drive**

Ambition and drive were developed throughout the year. The author normally would be more accustomed to using drive to achieve sporting objectives. Academic objectives require another type of drive to achieve. Sometimes in sport competing against another competitor would determine the drive required to achieve a goal. For example, when a player takes their position on a sports field of play they would assess their opponent’s motivation and their ability and then determine the difficulty of the task at hand. In some cases, the opposition may be too inadequate and require less drive to defeat or in other cases too strong where using too much drive would be considered a waste of energy. These examples can also be true of academic objectives too. However, this project required the author to branch into areas where no direct competition could be utilised. This lack of competition meant the author would need to learn to drive to become better ,not to better an opponent, but to better oneself and one’s ability.

**Be prepared for difficulties**

Being prepared for difficulties is the best piece of advice this author can provide for anybody in any walk of life. This was the most beneficial concept the author learned throughout the timeline of the project. This concept should not be mistaken for “what can go wrong, will go wrong.” The author learned that the concept of “what can go wrong, will go wrong” can be averted simply through being prepared to face difficulties. No matter what aspect of a project or recreational hobby a person is currently interacting with, there exists a possibility that the author will face a difficulty. This possibility could easily be described as an inevitability. A difficulty can be a problem of any size or a problem that carries with it somewhere between minimal to catastrophic repercussions. Being prepared for these difficulties arising can benefit a person in many ways. Initially, just being aware of the possibility of a difficulty arising will lessen the initial impact or panic associated with problems. This will hasten the time it takes to begin to formulate a solution to the problem. Furthermore the acknowledgement of the likelihood of difficulties when assessing the timeline to be put in place for the project will allow a person to a lot some time for difficulties and will thus provide a more accurate timeline and if working under a deadline will allow for a more accurate estimation of how far along the project can be worked to within the allotted deadline. The most important and what this author believes will become the most influential benefit of being prepared for difficulties lies within being mentally prepared to meet a problem along the journey to success, both within a project and within all other aspects of life itself. When a problem is met, people will deal with this problem better than others. Some will essentially shut down in a state of panic and other will relish the challenge and rise above it. Being prepared to meet a challenge will substantially increase the likelihood of relishing the challenge. Acceptance of the inevitability of a problem arising will drive a person forward to meet the challenge and over come it. This drive to be tested and to overcome challenges grew and grew substantially for this author as the project progressed. This drive got to the stage that the author would assess a task that needed to be completed and if it was not difficult or challenging enough the author would increase its difficulty by even sometimes seeking an issue the author did not know the answer to off the top of ones head and then proceeding to implement this task into the project. It is the authors belief that this drive that was found in the project was the major reason for the success of the project.

**Technical**

**Python**

Before the project the author had an acceptable understanding of python at it base form and the different data structures pythons base form provided. These include tuples, dictionaries and lists. Throughout the project the authors understanding of the inner workings of python increased greatly. The author would consider himself an expert in pythons two main data structures, dictionaries and lists. Due to the vast number of instances of dictionaries and lists the author was exposed to an enormous amount of issues and was forced to fix these issues himself through trial and error among other lesser used techniques. To fully communicate the ability the author has accrued with python would be the following statement, The author has completed the creation of two machine learning algorithms using python and no more than 5 libraries which were used only for some small issues and did not contribute a significant amount of work to the algorithms.

**Data Science**

The author had no knowledge of Data Science whatsoever before this project. The author now believes himself to be at an intermediate level of Data Science. One such demonstration of this ability comes in the form of a project the author was expected to undertake as part of the course the author is enrolled in. The author completed a project of significant quality and one that outstripped anything that was expected of the author based upon the knowledge gained from the course the author was enrolled in. The author has been informed that the authors knowledge of Data Science would rival that of a person that has completed a master’s in data science. Of course, data science covers the areas of machine learning and data mining, two areas that the author has made significant improvements in throughout the project.

**Algorithms and Data Structures**

This project required the author build a custom algorithm. The project also required the author to build a Q-Learning algorithm from scratch without the use of libraries. Both of these aspects expanded the authors knowledge of algorithms a massive amount. The author would now consider himself competent at both the designing of algorithms and the building or software engineering of algorithms too. Through the development of the q-learning algorithm the author was required to develop a linked list. This linked list as demonstrated in Fig.1 was extensive and more complex than any linked list the author was ever exposed to. This vastly increased the authors understanding of these data structures.

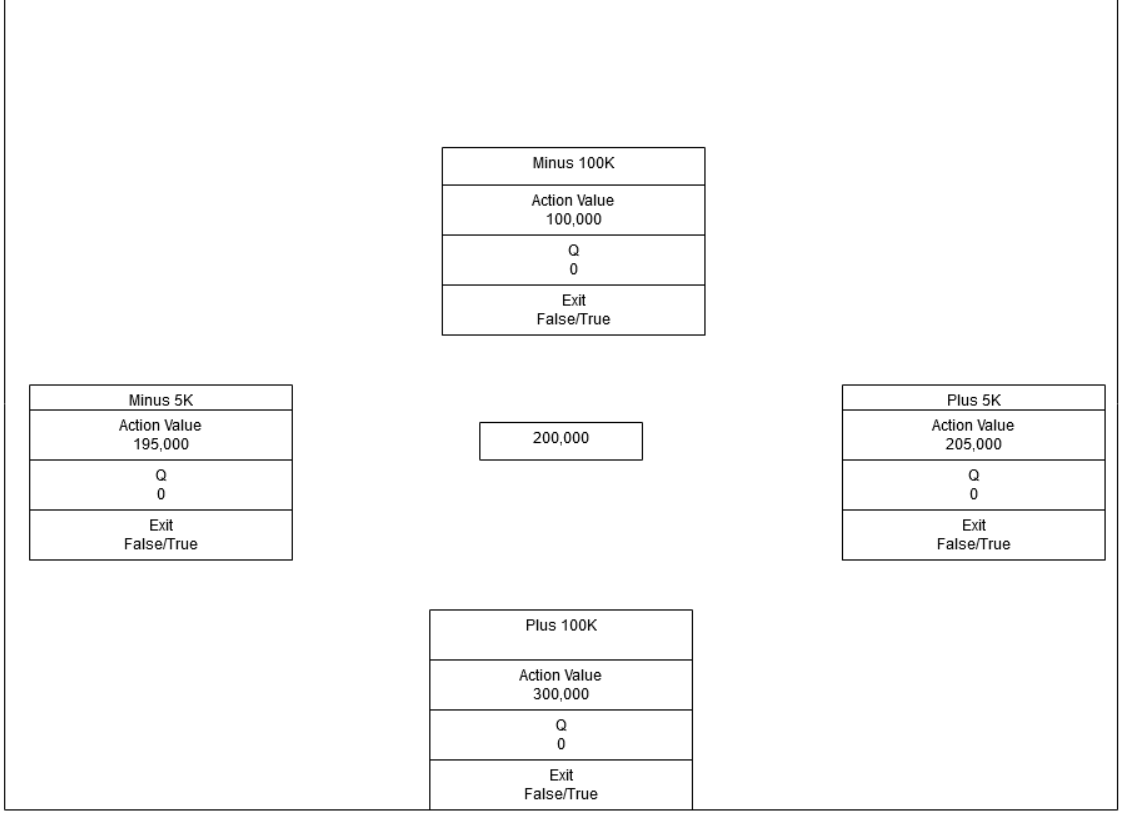


Fig.1

The figure above represents one node in the linked list. Each preceding node will contain the same structure and different values. The Action Value is the implementation of a pointer for a python dictionary.

**Review of project**

The review of the project will detail the progress of the project and then provide some key information about the project. This will include what went well and what didn’t among other aspects. The insightful information provided in this section will show the reader a more personal side to the projects process.

**What went well**

**The Crawler**

Overall the crawler can be considered a resounding success. As the research aspect of the project was proceeding and even producing some wonderful results there became an apparent need for a data set that could not be acquired without a crawler. The author had very little experience of a crawler or a scraper. Despite this the author was able to complete the crawler in the space of a week. On top of this obvious success the crawler never needed to be altered throughout the entirety of the project. This non-altering demonstrates the initial and lasting success wonderfully.

**The Research**

The research was another success. The research covered a number of areas. Algorithms, machine learning, data mining, data science, regression predictions, house price prediction and data science were all researched under their own banner. That is to say they were researched individually and not as a collective of another topic despite the obvious inclusion of some topics in other topics. The final research topic that was researched was coefficient of variation which was researched one week before the deadline. The amount researched, and the amount learned from the research proved to be substantial. Overall the research was interesting and enjoyable and a complete success.

**Time management**

Overall the management of time was exceptional by the author. Despite the clear and obvious over ambition the project actually exceeded the ambition and this was definitely in part due to the time management of the author. Although work ethic and further ambition could account for the success of the project the time management is the aspect that really made things possible and opened up opportunities at times that allowed for these opportunities to be grasped.

**What didn’t go well**

**Sickness**

Unfortunately, in February which is a critical period for the project the author fell ill for three weeks which prevented any work being completed. February is especially pivotal for the project as it allows the author ,of any project, to get a head start on the project and to achieve the building of the foundations of the project ahead of schedule. The illness of the author meant the author had to begin the third iteration of the project without the foundations in place. Normally this would be acceptable, however, due to the authors ambition this put the extent of the project’s success in jeopardy. If it wasn’t for the aforementioned work ethic, then the project would not have been completed to the extent that it was and this sickness is the reason the author needed to put in so much work.

**Testing**

Testing the project was the final task. Coefficient of variation was the chosen method of testing. This method was chosen due to the results of extensive research on how to test regression and more specifically how to test neural network regression. Neural network regression was used as a reference point due to the similarity of the logical make up of how results from q-learning and neural network regression would be similar to each other. Unfortunately, the method chosen was incompatible with q-learning. The author would also like to express caution when using coefficient of variance with neural networks despite the arguments that googling a test for neural network regression will provide. The author does not believe them to be compatible. Due to the late timeframe that the testing was undertaken there was no time for a new method to be produced and as such the testing was incomplete.

**What would you change**

**Testing**

As mentioned above the author would change the testing aspect of the project. There has not been sufficient time to research an alternative but the author believes that several testing exercises would be more beneficial than just a singular test.

**Nothing Else**

As the project stands at this time in its current state, the author is content with its makeup. There is nothing within the project or its structure that the author would alter. With regards to the process of developing the project the author cannot prevent sickness so this is not a viable change and as everything else outside the technical aspects went almost without incident the author has no changes to suggest.

**What future work can be done**

**All Ireland Implementation**

The project can be implemented for the entire of Ireland. This is currently possible and would require minimal work but would require substantial computing power and time.

**Future Price Prediction**

The price currently predicted is the current selling price of the property. The data retrieved and cleaned could be easily utilised to predict future changes in price in each area throughout Ireland. This new predictions could be used to pinpoint areas that would be considered to be fast growing in property prices. An investor could then use this information to make some capital. Predicting the future prices and subsequently the growth of each area would be very beneficial from a monetary perspective.

**Refinement of the Custom Classification Algorithm**

While the author is exceptionally happy with the end result of the classification algorithm, as with any project there will always be room for improvement. To improve the classification algorithm, one would require a lot of hours of micro management. This process would be tedious and frustrating but could yield some excellent results and could provide the best return on any potential further work due to it capabilities of being implemented anywhere in the world.