

Property Price Prediction

Alan Cahill

C00174533

Lecturer:

Dr. Lei Shi

Research Report

**Abstract**

The purpose of this document is to present the research that has been undertaken for this project. The research topic will be explained along with the reason for undertaking the research. Each topic covered was researched for the purpose of implementing said topic within the project. At the end of the research the document will state the decisions on which topic or technology was chosen for use within the project. The reasons for the choices will also be discussed.

**Table of Contents**

**Introduction5**

**The Property Market6**

Transactions**6**

Valuation and Details**6**

Sellers**7**

**Property Price Prediction8**

Valuation Prediction**8**

Valuation Prediction: Kaggle**8**

Future Price Prediction**8**

**Stock Market Price Prediction Using Machine Learning9**

**Irish Property Market Data Availability10**

Residential Property Price Register**10**

Online Property Advertisement Websites**10**

Daft.ie**10**

**Data Science12**

Data Science**12**

Data Pre-Processing**12**

Data Analytics**13**

**Machine Learning15**

Supervised**15**

Un-Supervised**15**

Reinforcement Learning**15**

Machine Learning Algorithms**15**

Logistics Regression**15**

S.V.M**16**

Multiple Regression**16**

Decision Trees**19**

Random Forests**19**

Naive Bayes**20**

Linear Regression**20**

Neural Networks**21**

**Spiders23**

**Q-Learning**2**4**

**Machine Learning25**

**Clustering26**

Density Based Spatial Clustering on Applications with Noise**26**

K-Means Clustering**27**

**Graph Theory28**

**Technologies29**

General Use Technologies**29**

Spider Technologies**33**

Data Storage Technologies**34**

Data Pre-Processing Technologies**35**

Clustering/Classification technologies**35**

Data Visualisation and Statistics Technologies**35**

**Literature Review37**

Real estate price forecasting based on SVM optimised by PSO:**37**

Housing price prediction using neural networks:**37**

Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data: **38**

Ensemble ANNs-PSO-GA Approach for Day-ahead Stock Exchange Prices Forecasting:**38**

**Decisions Made39**

Project Structure Decisions**39**

Technology Decisions**41**

**Bibliography43**

**Introduction**

The purpose of this document is to provide an overview and a description of all the research undertaken to aid the completion of this final year project. The aim of this project is to create a computer program that can make predictions of house prices based off the historical data of the housing market.

The process involved in achieving this project’s aim contains several subsections and subprocesses. The process itself is as follows, retrieve a data set, analyse the data in this data set and then using Data Science techniques prepare the data set for use with machine learning models. Once preparation is achieved the data will be used to train a Q-Learning algorithm. Also, after further analysis of the data and the fundamentals of the Irish property market it was decided that a classifying algorithm was needed to classify sub areas in each specific town or area. This algorithm required innovative creation and as such has its own subsections.

The areas covered in this document are as follows:

* The property market
* Property price prediction
* Stock market prediction using machine learning
* Irish property market data availability
* Data science
* Spiders
* Q-Learning
* Clustering
* Graph theory
* Techniques
* Technologies
* Literature Review

Each of the areas mentioned here will have their own subsection. These subsections will be detailed during the discussion of each area.

At the end of this document decisions will be made and will be conveyed with a discussion on the reasons these decisions were made. Throughout the progress of the project these decisions may change, and these original decisions will be described along with a discussion as to the reasons the decisions were altered.

**The Property Market**

Focusing on the core fundamentals of the property market research on this topic was very brief in comparison to other areas mentioned in the introduction. Very simply broken down the property market contains buyers, sellers, proprietors and property auctioneers. Of these entities’ buyers, sellers and property auctioneers are considered active in creating transactions in the market. These transactions are the fundamental basis that this project will be constructed on. It is also necessary to state that Ireland will be the focus of the project and furthermore the focus of the research.

**Transactions:**

Almost all the transactions within the property market, and a high enough percentage for this project to assume a totality, contain a buyer, a seller and a property auctioneer. To begin these transactions a seller must approach and employ an auctioneer. The auctioneer will then undertake the sale of the seller’s property. To achieve a sale the auctioneer requires several things. The fundamental of what is needed are a property’s valuation and details, access to said property, the required legal documentation and finally a seller. For the purpose of this project there will be no need to research the documentation and there will be no requirements for property access.

**Valuation and Details:**

The details of a property are fundamental to assessing the value of the property. These details are usually collected manually by the auctioneers or in some instances they are conveyed to the auctioneer by the seller of the property. Using these details, the property auctioneer assesses the property and from their professional judgement they ascertain a valuation for the property. This valuation will be conveyed to the seller and from further discussion, while this valuation will not change there may be a decision made to sell the property at a lower price, for any amount of reason too broad to mention here, or for the same broad spectrum of reasons the decision may be to attempt to sell the property at a higher valuation. These interactions between the seller and the property auctioneer are not public record and this is an important point that will be discussed again later.

To explain how to achieve access to these details and valuation the topic of sellers must be discussed first.

**Sellers:**

In order for a property to sell the auctioneer must of course have a seller to sell the property to. The auctioneer in some situations will have a seller already earmarked for the property in question. In these instances, the negotiations will begin and in a successful instance there will be a sale completed with the authorization of all parties involved. This transaction is then declared to the Revenue Commissioners and all other proper legal requirements fulfilled. The important aspect here is the details of the sale become public knowledge. These details are published on the Residential Property Price Register. The details published here are the price date and address as per the details shared with the Revenue Commissioners. This public information can be access and will be discussed later. This is the process in Ireland as mentioned previously.

In other situations, the property auctioneer will not have a buyer earmarked and must search for someone to become the buyer they require for the sale. There are of course several ways to achieve this. Word of mouth, placing local advertisements in local media, the use of a property as an office to allow people to visit are all viable and widely used methods for finding sellers. On a side note word of mouth would be more widely used to give loyal customers first option or a head start on opposition buyers. Despite the use of these methods none of them provide this project with any valuable or useful assets. The final method that will be mentioned will provide the project with the core asset for the project. This asset is obviously the data set to be used in the project and the method to be discussed is the use of internet-based websites that allow the public to view properties and their details in their own home. This method has reduced the workload for auctioneers but most importantly it allows the developer of this project to get access to details of thousands of properties throughout Ireland. These websites will be discussed and detailed later.

**Property Price Prediction**

The research for property price prediction was approached with a naïve stance. What this means is there was no previous knowledge of property price prediction and as such there was no pre-determined route or final destination. To summarise everything learned was new knowledge with no indication of importance and without a destination in mind the researcher was required to research whatever aspects appeared throughout the research. The first aspect that was appearing regularly which was then pursued as a specified sub topic was using the data of a property to predict the valuation this property would receive.

**Valuation Prediction:**

There is an abundance of previous work on the topic of valuation prediction. There are research papers published and there is an array of projects on GitHub that would be classified under this topic. Furthermore, there are even competitions being run by Kaggle to ascertain the best program for predicting the valuation of a property based off of its details.

**Valuation Prediction – Kaggle:**

Kaggle[7] are currently running a competition of their website that challenges the contestants to predict the final price of a home in Ames, Iowa. Kaggle provide the contestants with a data set of houses that contains 79 explanatory details of the properties. They recommend in the description to use [1] ‘advanced regression techniques like random forests and gradient boosting.’ The current leader of the competition has a score of 0.000 based off of the Root Mean Squared Logarithmic Error of their algorithm. Most of the projects viewed by the researcher utilised the Python library scikit-learn. This is not the direction that this project will be undertaking.

**Future Price Prediction:**

Researching the prediction of the future price of a property proved to be a difficult task. While there are an abundance of resources for predicting the property market there are no resources that this researcher could find for predicting the value of individual properties in the future. There are several research, media and internet-based articles detailing the prediction of the property market environment. These articles detail how a person can or has been able to predict the direction that general values in property will fluctuate in the future based off factors that affect this fluctuation. One such factor is the level of interest rates. This approach whilst proved effective is not applicable to this project. The reason for this is this project will be undertaking the prediction of prices based exclusively off historical data of the properties in the property market.

**Stock market price prediction using machine learning**

This topic was researched to use it as a reference point for predicting the future price of property. After the results of the research for predicting the future price of a property were conveyed to the mentoring lecturer the lecturer recommended this area as a reference point. Something that could be predicted about this area is the abundance of material available. This area is extremely profitable. Googles stock prices on August 20, 2004 were valued at 54.16 USD. This has risen throughout the years to reach a high point on August 3, 2018 at 1238.16 USD and as of March 1, 2019 are trading at 1148.52 USD. Due to examples like this there is a multitude of research and projects for predicting stock price and within these projects and research there are some that concentrate on using historical data. Fundamentally the research into stock market price prediction proved to the author that it is possible to predict prices off of historical data while also producing some reference points to use a starting point.

**Irish property market data availability**

The previous research areas offered up a varied list of questions and answers. However, at the time of researching there was no data set available or earmarked for use for this project. The previous research all detailed the dependence of this project on the data set that will be used. A search for a dataset was begun and the results proved at times to be difficult and challenging.

**Residential Property Price Register:**

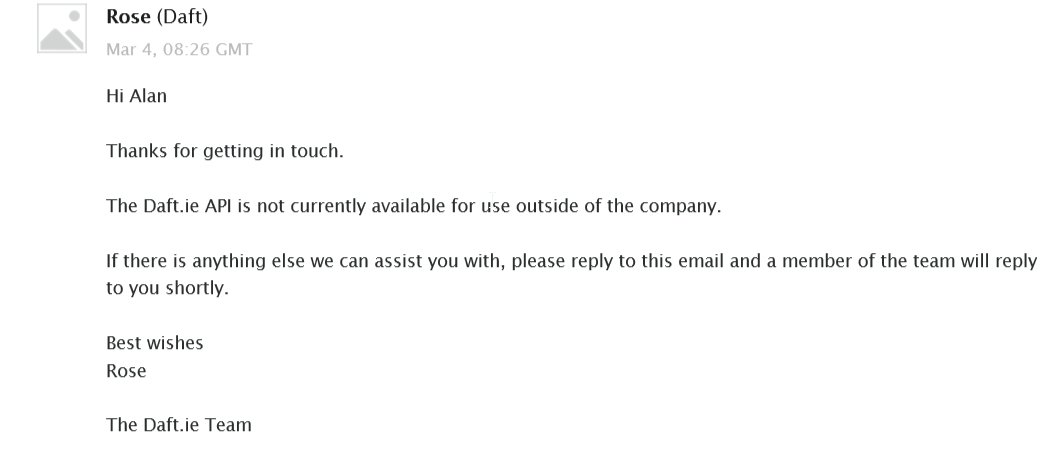
The Residential Property Price Register as detail before has a data set available describing the transactions that are declared with the Revenue Commissioners in Ireland. This data sets contains the address, sale price and date of sale of each of its entries. This is unfortunately where the details end. With no property details it would be impossible to determine any kind of predictions on house price. However, this dataset will become very valuable in the future by detailing the actual sales price of a property, removing any discrepancies of human error or opinion.

**Online Property Advertisement Websites:**

Online property advertisement websites were the only source of data that was found other than the residential property price register. These websites contain any property that is advertised on their website. The properties in question are primarily advertised by property auctioneers at a cost given to the website in question. There is a plethora of websites that provide this service for property in Ireland. These include but are not limited to Property.ie, Daft.ie and MyHome.ie. Trovit.ie also provides this service but only for rental property. Without exploring into these websites with too much details they all essentially provide the exact same service with some minor and cosmetic differences. With some additional research that entailed comparing the amount of properties in each particular area on each website and some research on the publicly most popular website it was decided to use Daft.ie to gain a dataset. Daft appeared to contain more properties and was more widely known to the public the researcher interacted with.

**Daft.ie:**

Daft.ie[8] contains an API to access the data on its website. With some implementation of custom-built code this API would provide a dataset that the project could utilise. Unfortunately, after contacting Daft.ie they stated that the API was only accessible to developers inside the company. This information is detailed in Figure 1.1. Unfortunately, the original e-mail was lost so the researcher contacted the website again to glean a new reference e-mail. Once this e-mail was received it was clear that a spider would be required to traverse the website and glean a dataset. The data gleaned from the website will be described later.

****

**Fig. 1.1**

**Data Science [9]**

This project is a data science project. As such data science needed to be researched and learned to a level that allowed this project to be completed. The reason for stating that data science would only be researched and learned to a level that allowed this project to be completed is because data science is such a vast topic that researching un-necessary aspects of it would jeopardise the project’s completion. Despite this statement the vast quantity involved in researching data science was often strenuous and required a large allocation of time. Data science was a new discipline for the researcher so anything that was researched was also a new discovery or a new discipline learned. Therefore, anything described under the following areas is what was learned and what was discovered along with being the description of the topic itself.

**Data Science:**

Data Science is an interdisciplinary field that uses scientific techniques, methods and algorithms to derive knowledge and insight from data. Data science is the basis of this project, providing the researcher with ample reason to research it. Data science is a combination of several aspects. The use of machine learning, data analytics and data preparation are the essential aspects of data science which will be utilised in this project.

With such a vast collection of areas within each aspect they will broken down into subsections. The first being data pre-processing, the second being data analytics and thirdly machine learning. Data analytics and machine learning are hugely vast areas themselves which will require the author to again supplement each area with sub-sections. For this reason, machine learning has been separated into an individual topic. Data Analytics whilst being broad was not researched with enough detail to require an individual heading.

**Data Pre-processing:**

Data pre-processing itself is the area that many believe to be the key area of data science. It is widely stated that any data science project will spend anything close to eighty percent of its time on data pre-processing. Any data set that has been gleaned from anywhere no matter the contents or the reliability of the source will require some data pre-processing. Some aspects involved in pre-processing were cleaning, integration, transformation, reduction and discretion. Due to the high workload the researcher did not initial research on these topics to instead focus on more difficult and pressing matters. As the project continued and some pre-processing was required the topics required were researched. These topics researched are cleaning, integration and transformation.

Cleaning:

The area of cleaning includes some sub-topics. Handling of missing data, resolving inconsistencies and removing outliers. Handling rows with missing data can be handled by addition of data to these rows or by removing these entries in their entirety from the data set. Resolving inconsistencies is a much broader area that will depend exclusively on the data set in question. One such example is a dataset that has numbers stored as a string but in some entries the area that the number is supposed to be there is a word. This problem would have to be solved based upon the requirements of the task at hand. In some cases, the number will be required and therefore the rows with words may be removed or a replacement number inserted whereas in another case the word would be required resulting in the previous solution being reversed. Any inconsistencies in data have no global solution and must be solved on an individual basis. Of course, this is true with all cleaning aspects, but it is a very prominent problem with inconsistencies. Outliers are data that sit outside the normal area that the set lies within. These can be simply removed but must be found to remove them. An example of an outlier is a number that lies an extreme distance away from the mean of the numbers in the set. Removing these outliers will allow the future model to arrive at more accurate results.

Integration:

Data sets come in many shapes and forms. Data sets have no global or distinct form of containment. As such for a dataset to be portable with some programs such as Rapid Miner or even some privately built implementations of machine learning the dataset will sometimes need to be converted to a certain type of data. Examples of such are, an SQL table, Mongo DB document, CSV and JSON. This is data integration, and whichever is suitable for the project is the format the data should be transformed into.

**Data Analytics:**

Data analytics is the discipline of analysing data to determine a pattern within the data or to extract rich information from it. Data analytics tends to utilise descriptive or inferential statistics, probability distribution, to achieve its ultimate goals. The following areas of data analytics were researched in detail for utilisation in this project:

* Statistics
* Probability
* Hypothesis and Inference

These areas will now be explained.

Statistics:

Statistics is a branch of mathematics dealing with data collection, organisation, analysis, interpretation and presentation. When using statistics, it is conventional to have a statistic model to study. Statistics is used throughout Data Science and is used to prepare the data for machine learning. This preparation is achieved through techniques like ascertaining key information from the data and finding correlation among other techniques too broad in variety and bulk to mention here.

Probability:

Probability measures the likelihood that an event will occur. Primarily in data science probability is needed to understand and utilise machine learning techniques. Probability will also be used to ascertain the likelihood that a model will return results containing errors.

Hypothesis and Inference:

Hypotheses and inference are built around probability. A hypothesis is a supposition or proposed explanation on the basis of limited evidence as a starting point for further investigation. When one has a hypothesis, one will test the truthfulness of the hypothesis using testing techniques. When one does not have a hypothesis, one will infer a hypothesis using inference techniques.

**Machine Learning**

[3] “Machine-learning algorithms use statistics to find patterns in massive\* amounts of data. And data, here, encompasses a lot of things—numbers, words, images, clicks, what have you. “By utilising machine learning computers can begin to act without explicit programming. Instead of explicit instructions the computer uses patterns and inference to act. Machine learning is commonly referred to as Artificial Intelligence. This is however incorrect. Machine learning is merely a subset inside of artificial intelligence. Machine learning is another vast area that has been researched and because of this the explanation of machine learning must be structured into manageable sections. The first breakdown will be the three types of machine learning:

* Supervised
* Un-Supervised
* Reinforcement Learning

**Supervised:**

A supervised machine learning algorithm will utilise data that has a label on it to ascertain a pattern in the data. The machine learning algorithm will asses the data and by referencing the label will decide upon a formula for labelling future data objects. When the algorithm is used on a new un-labelled data object it will assess the data and then it will utilise its formula for labelling new data.

**Un-Supervised:**

An unsupervised machine learning algorithm will utilise data that has no label on it and will instead categorise data objects based on what it believes is correlation. By assessing the data and its attributes the machine learning algorithm will find new correlations using its own methods. This correlation will be used to categorise the data. New data will be categorised by assessing its correlation to the previous data. In some cases, the new data will be entered into the model and then the entire model will be freshly categorised.

**Reinforcement Learning:**

Reinforcement Learning for want of a better explanation utilises trial and error. The algorithm will attempt something and will then record its success or failure before attempting something else. The algorithm is essentially trained by rewarding and disciplining the algorithm. A reinforcement learning algorithm differs from tradition machine learning algorithm by its non-assumption of complete knowledge of an exact mathematical model. A further explanation of this is that the algorithm does not assume that it knows everything about its environment (the data’s model) and instead assumes that it only knows a partial segment of the information on its environment. The algorithm acts in accordance of this partial knowledge through continuous exploration of its environment. To achieve all this reinforcement learning is usually modelled as a markov decision process. A brief explanation of a markov decision process is one which states that a markov decision process contains four variables. S, a set of states. A, a set of actions. Pa, a probability of state transition after action a. Ra, the reward for taking action a. Calculating and utilising these variables is the basis of a markov decision tree and the foundation of a reinforcement learning algorithm.

The following image Fig. 1 is a visual representation of how a reinforcement learning algorithm would operate.

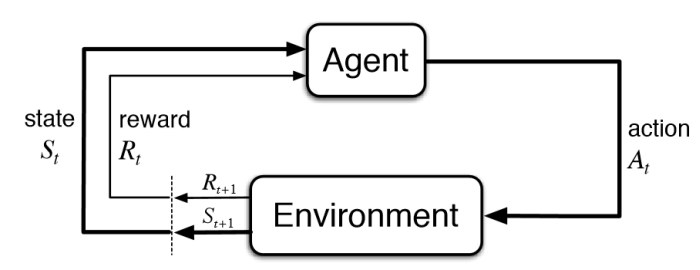


Fig. 2

***Machine Learning Algorithms:***

There are several utilised machine learning algorithms used throughout research and in industry today. When learning about machine learning there was a need to research various machine learning algorithms. Some of these algorithms were research in more depth for a proposed use in this project before a final decision on one algorithm was decided on. Furthermore, when it was decided that a clustering algorithm was needed clustering and some of its algorithms were researched in further detail. These algorithms will be discussed below in order beginning with the least researched algorithm and excluding Q-Learning and Clustering. The algorithms to be discussed are:

* Logistic Regression
* Support Vector Machines
* Multiple Regression
* Decision Trees
* Random Forests
* Naïve Bayes
* Linear Regression
* Neural Networks

**Logistic Regression:**

Logistic regression[10] is a type of regression that is used as a prediction mechanism. Normally when using logistic regression, the variable being predicted will be binary i.e. will have only two possible outcomes, usually pass or fail, win or lose etc. The line for logistic regression is plotted on an x and y axis graph and takes a form that resembles S as demonstrated in Fig.2. This is called the logistic function and is found using the formula in fig.3.

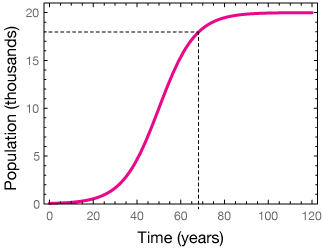


Fig.2 Logistic Regression Diagram



Fig.3 Logistic Function formula

**Support Vector Machines:**

SVM[11] is a supervised machine learning technique that can be used for regression and for classification. ‘The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.’ This is achieved by finding the hyperplane that will separate the points of each class by the maximum distance. The reason for choosing this particular hyperplane is to maximise the confidence when making predictions in the future. An example of finding the hyperplane with the largest distance on a 2-dimensional graph is shown in Fig.4

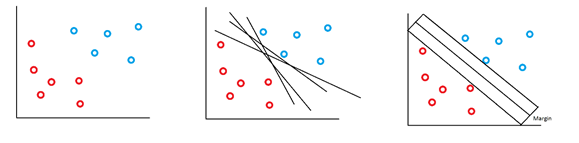


Fig.4 Support Vector Machine on 2-Dimensional Graph

**Multiple Regression:**

Multiple regression is an extension of linear regression. This is used when it is necessary to predict the value of a variable based off the value of two or more other variables. A multiple regression model will allow the relative contribution of each of the variables on the variable being predicted to be determined. It will also determine the overall effect of the variables together on the variable to be predicted. Multiple regression techniques will in general estimate a linear equation in the form, Y = a + b1\*X1 + b2\*X2 + ... + bp\*Xp. In Fig.5 there is a simple multiple regression demonstrated but in Fig.6 the extent that a multiple regression can grow to is visualised.

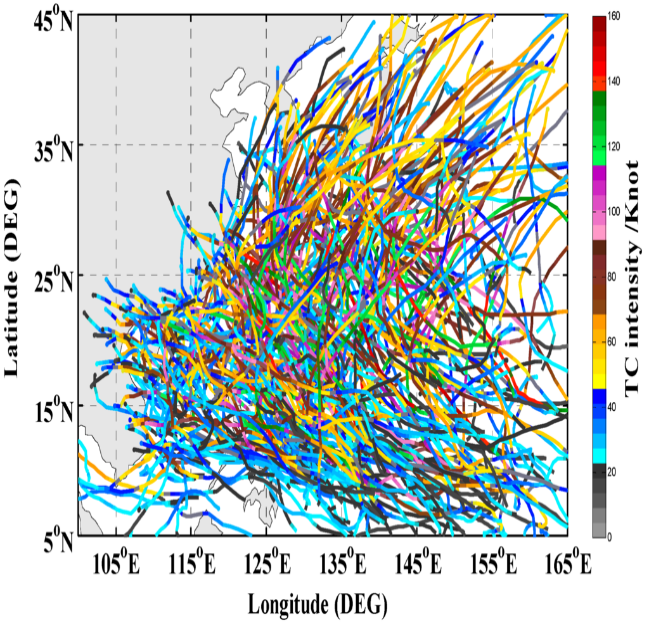
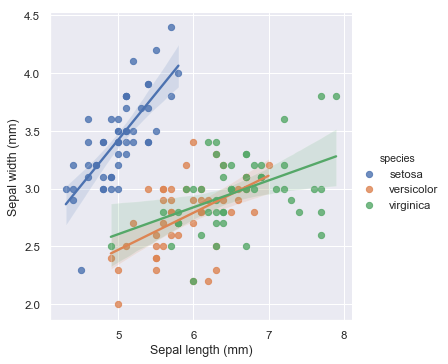


Fig.5 Simple multiple Regression Fig.6 Complex Multiple Regression

**Decision Trees:**

Decision trees is a supervised technique that is normally used for classification. Through the use of entropy, the data set is broken down continually by splitting the data set based off the information that can be gained from a split. The higher the entropy the more information can be gained. If entropy is high for a split this means that this split can be used a question in the decision tree. Each split or decision creates branches which are continually broken down until all data objects on that side of the tree are categorised or there are no more splits available. When a new object is put through the model, the model will use its previously made decisions on the full dataset to classify the new object. The formula for entropy is given in Fig.7. Fig.8 is a visualisation of a decision tree.

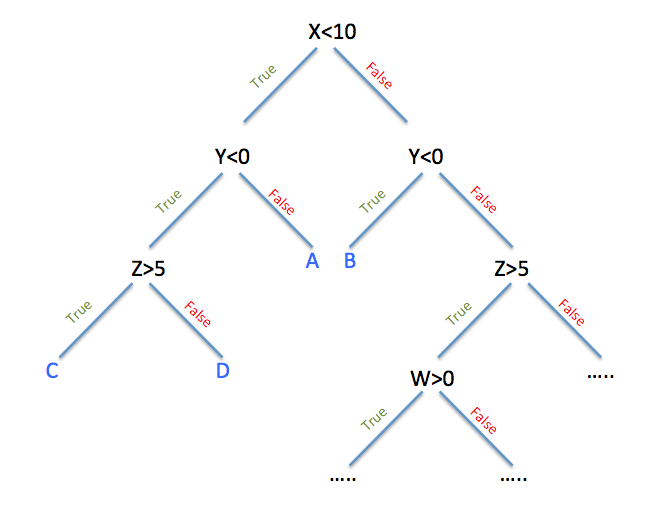


Fig.7 Entropy Formula Fig.8 Decision Tree

**Random Forests:**

Random forests[12] employ the use of decision trees. To create a random forest on must use several decision trees. Each decision tree is run through the data and the results are found. When deciding on a final decision the mode of the results is used for classification and the mean is used when using random forests for regression. Essentially random forests are the use of several decision trees with the final result for classification the result that appears most often and the final result for regression being the average of the results of the decision trees. Fig.9 shows a random forest simplified.



Fig.8 Simple Random Forest

**Naïve Bayes:**

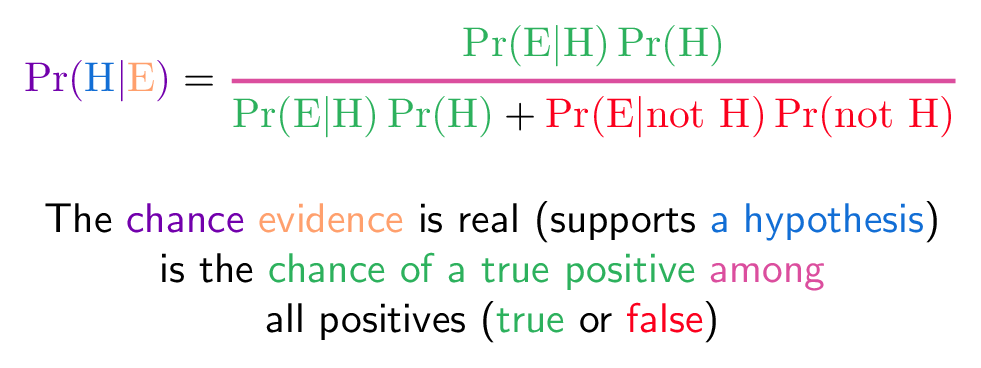
Naïve Bayes is a supervised algorithm that utilises Bayes theorem. Naïve Bayes uses probability to predict the likelihood of an event based off of prior knowledge of the event. Usually used for classification the Naïve Bayes algorithm will present its predictions of what class a variable should be based by referencing the highest probability it finds for that variable on each of the classifications. Fig.10 is the formula for Bayes Theorem

Fig.9 Bayes Theorem

**Linear Regression:**

When plotting the relationship between a variable and one of its descriptive variable the process becomes know as linear regression. If X is the descriptive variable and Y is the dependant variable the equation of the linear regression line is: Y = a + bX. Image. 11 Demonstrates linear regression.

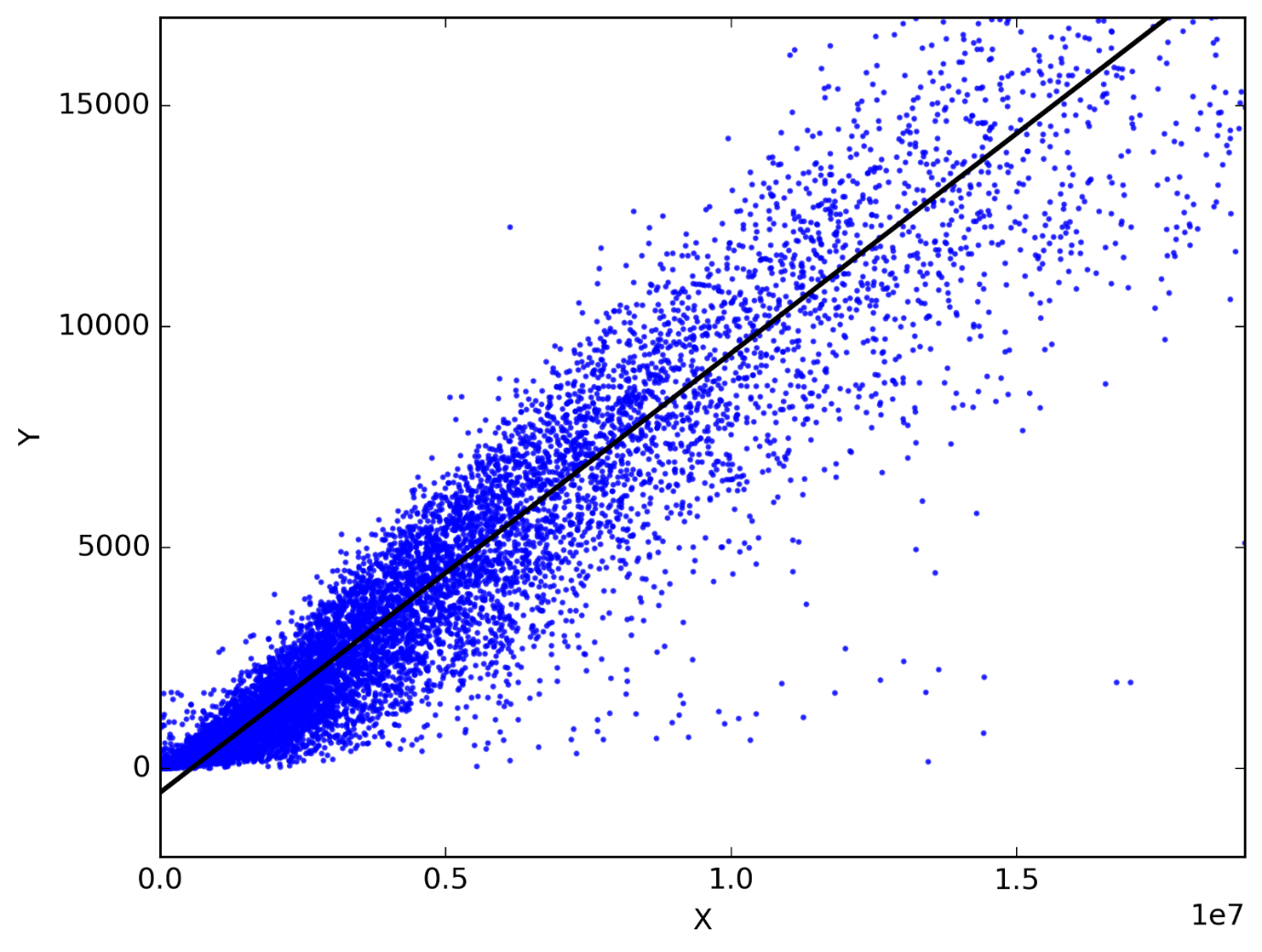


Image.10 Linear Regression

**Neural Networks:**

To understand neural networks[13], one must delve into biology which is where neural networks originated. In the human brain there exists a phenomenon called a Neural Circuit. A neural circuit is a population of neuron that are connected by synapses. A synapse is a biological tool used to by neurons to pass electrical signals. Fig. 12 shows a depiction of a neural circuit in the brain.

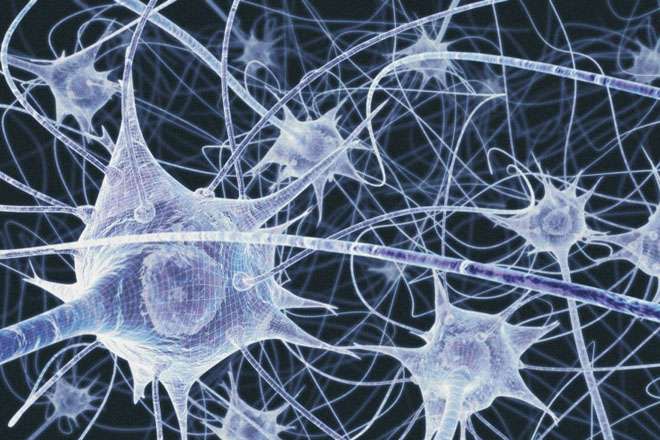


Fig. 11 Neural Circuit Depiction

From these neural circuit’s scientists created an artificial duplicate called Artificial Neural Networks. Like the circuits the new ANN’s send messages from one point to another. When the message is received some calculations are undertaken before a message then sent forward to the next position or neuron. To train a neural network a method called back propagation is used. By referencing a data set with already predefined answers the calculations are modified starting from the one closes to the answer moving backwards to the beginning. The object is sent through the neural network again until it reaches the end where if necessary the back-propagation sequence is repeated until the answer found at the end matches the predefined answer. Fig. 13 demonstrates a neural network being used to classify an animal as a cat or dog.

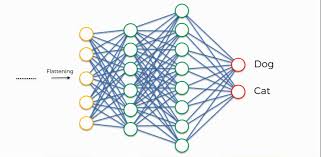


Fig. 12 Neural network

**Spiders**

A spider is a program that is built with the intent to explore the world wide web. Mostly a spider will be programmed to explore only an individual website. The spider will be given a URL link to visit. When it visits this web page it will then as per instruction retrieve the relevant links on this webpage. Not all links will be retrieved, just the links the instructions specify. These links will be added to the list of links to be visited which is traversed until there are no links left. As the spider traverses the links it continues to retrieve the new links. Sometimes the spider will retrieve data as it traverses the links. It utilises a scraping mechanism to achieve this. The scraping mechanism will retrieve the specified data from each of the links the spider brings it to.

**Q-Learning**

Q-Learning[14]&[15] is a reinforcement learning algorithm. It does not require a model of the environment which makes q-learning a model free algorithm. For most algorithms to succeed a data model is needed. In most cases this data model will need to be complete. The algorithm will then use this data model to make itself a decision-making system. Q-learning instead of utilising its knowledge of the model will use trial and error to make decisions. Despite this small difference between q-learning and other reinforcement algorithms the rest of what was mentioned in the section on reinforcement learning rings true here. The process and mentioned is the same here. Other versions of q-learning include deep q-learning, double q-learning and double deep q-learning. The algorithm for utilising the reinforcement learning process in q-learning is as follows in Fig. 14:

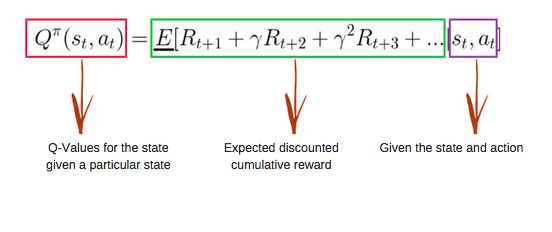


Fig.13 Q-Learning formula

To utilise Q-Learning one must initialise an environment. This environment is one that allows for decision making. Each position in the environment will allow for decisions to be made, contain a name or position identifier and a current state. Moving from one position to the other(action a) will change the sate of the original position based upon the algorithm (reward) and the state of the position being moved to. Fig. 15 demonstrates a Q-learning environment while Fig. 16 demonstrates how Q-learning operates.

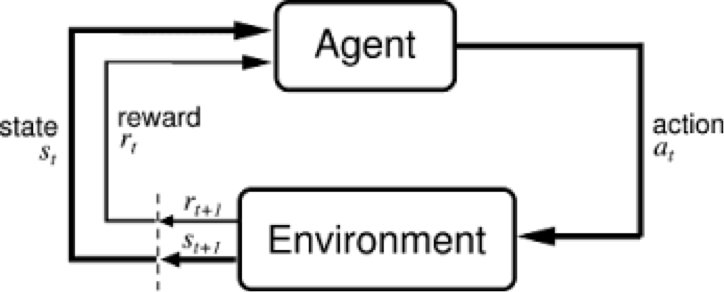
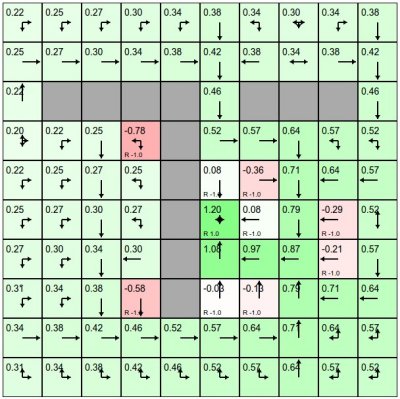


Fig. 14 Q-Learning Environment Fig.15 Q-Learning Operation Diagram

**Clustering**

Clustering was researched in depth after the discovery that the area of a property vastly effects the price of the property. So much so that a county in Ireland must be broken up into its individual towns or areas and then these towns or areas must be broken down further. With no official way of segregating the houses in the towns or areas an algorithm was needed to achieve this. Clustering was identified as the answer and was then researched.

Clustering is born of the idea that similar data objects will group together in some way and form clusters. Clustering in the machine learning sense is the creation or finding of these cluster. Clustering is an un-supervised algorithm. The algorithm will be provided with attributes to cluster the data with. Most commonly the number of clusters will also be provided. Once the algorithm has found the clusters it was asked to find it will not label them. This operation is down to the user, by looking at the underlying data the user will label the clusters. There are many types of clustering algorithms to choose from. The two main algorithms identified for this project were:

* Density Based Spatial Clustering of Applications with Noise (DBSCAN)
* K-Means Clustering

**Density – Based Spatial Clustering of Applications with Noise:**

DBSCAN begins with a solitary point. This point is marked as visited. The algorithm then calculates a neighbourhood. If a point is within the epsilon distance, then it is assigned as being part of the neighbourhood. If there are enough points then the neighbourhood is started at the current point, otherwise the point is considered noise. This neighbour hooding process is repeated for each of the neighbouring nodes identified in the first step. This is repeated until all the points in the neighbourhood have been visited and labelled. Once a cluster is finished a new unvisited point is found and processed. If there is a viable neighbourhood at this point it is generated. This process repeats itself until all the points have been assigned to clusters or marked as being noise. Fig. 17 demonstrates DBSCAN.

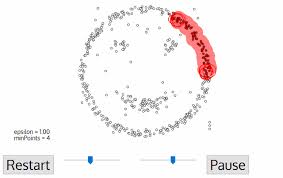


Fig.16 DBSCAN in the process of building a neighbourhood.

**K-means Clustering:**

For k-means clustering a desired number of clusters n is decided upon. Then an n amount of points are chosen on the graph and are used as starting points for the algorithm and will be the centre points of the clusters. All nodes are analysed and are assigned to the centre point that they are closest to. Once this has been achieved the centre of the cluster is calculated and the centre node is then moved to this point. This changes the makeup of the graph and which point belongs to which cluster. To solve this the points are analysed and assigned to the new clusters. The centre point is calculated and re-assigned again. This process continues until the centre points become static points. This final state is the results of k-means clustering. Any new point that requires clustering is added to the data and the k-means algorithm is ran again from scratch. Fig. 18 demonstrates K-Means Clustering.

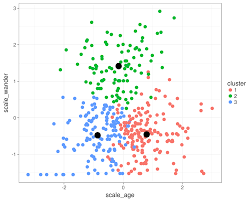


Fig.17 K-Means Clustering

**Graph Theory**

Graph theory is the study of graphs in mathematics. A graph is a mathematical structure composed of vertices, nodes or points which are connected by edges or lines. The study of these structures is graph theory. Specifically, the area researched involved the plotting of lines on a graph when given a point and the angle and direction of the line to be plotted. Furthermore, testing to see if a point was on a line was researched. The following formulae were discovered for use:

*y* = *mx* + *b*

This formula is the linear equation of a line in slope-intercept form.

d(P, Q) = √ (x2 − x1)2 + (y2 − y1)2

This formula finds the distance between two points P and Q

**Technologies**

This area will state and discuss the technologies that were researched in accordance with this project. These technologies fall under sub-headings due to their usage within this project. The libraries from the various languages will be discussed in the sub-sections.

These sub-headings are:

* General use
* The spider
* Data storage
* Data pre-processing
* Clustering/classification algorithm
* Data visualisation and statistics
* Q-learning algorithm

**General use technologies:**

The technologies described in this sub-section will be used throughout the project and under multiple or all the other mentioned sub-sections.

**Windows:**

Windows was chosen as the operating system the project would be developed on primarily due to the failure of the developer’s computer to allow a duel boot of windows and linux. When installing a version of linux mint the computer would continually freeze. For this reason, windows was chosen and linux was not researched. All further technologies were researched with the intentions of running them on a windows machine.

**Python:**

Python[5] will be used as the primary language for this project. Python is a high-level interpreted general usage language that emphasises easy to read code through use of significant whitespace. It was released in 1991 and has become very popular due to ease of use. Currently it is the most popular language used by data scientists globally. Along with the benefits of readability python also provides a multitude of libraries to aid the development of software with over twenty thousand available. Python libraries that were researched will be discussed later in the document. Python was chosen due to the developer’s prior experience with the language, its relevance to data science and its superiority in handling data when compared with the other languages that the developer is competent with. Those languages are Java and C++. These were not researched due to the early decision to use Python. R was not researched for the purpose of using as a primary language for this project but will be discussed later in a different area.

**PyCharm:**

PyCharm[6] is and IDE specific to the development of python. The developer had previous experience with PyCharm and with no other experience with a different IDE it was decided to use PyCharm as the IDE for the python aspects of the project. PyCharm also provides portability with Git and GitHub.

**Anaconda:**

Anaconda[16] is an open source distribution of Python. Anaconda provides its own command prompt, package management system and install (conda) and provides an interpreter. Anaconda comes with pre-installed libraries aimed at the scientific aspects of programming. The libraries included are ideal for Data Science and machine learning. It was decided to use Anaconda with each of the details mentioned providing part of the motivation for use. Unfortunately, due to some technical difficulties while developing the Anaconda distribution had to be removed from the project and in its place the Python interpreter which comes standard with Python 3.x was used.

**Python Interpreter and pip:**

When Anaconda failed during usage it was removed and the python interpreter that comes with the distribution of Python 3.x was chosen as its replacement. Instead of using conda for downloading and managing libraries for Python the Python interpreter can be updated and added to using pip. While it sounds much better to use conda for everything involved in package management using pip is a much quicker way to install and manage the packages. Pip does not need to resolve an environment, instead simply downloading the packages and then installing them. As a result, pip turns out to be incalculably faster than conda.

**Git:**

Git[17] is a version control system that allows tracking of source code in software development. Git is relatively easy to use and the developer had some prior experience with it. Once installed Git is accessed through a Git Bash, Git GUI or can be installed directly onto the command prompt.

**GitHub:**

GitHub[18] is an online repository for source code. It is directly portable with Git and like git it allows for tracking of source code. When a project is created on GitHub or Git it can be linked with the other and then using Git the project can be kept up to date on both technologies. Git Hub was chosen to store and backup code.

**DB Browser for SQLite:**

Db Browser for SQLite[19] is exactly that. A program that allows the user to browse through the data for an SQLite library. This appeared to be the easiest such program to install and use and as such as chosen for that reason. The initial need for this was due to the use of SQLite. With the data saved in SQLite there was no way to view the data without a specific program. DB Browser for SQLite was utilised for viewing the data throughout the development of the spider.

**Scikitlearn:**

Scikit-learn[20] is a python library that contains machine learning algorithms. The library has its own inbuilt functionalities for developing working machine learning models. Scikit-learn has functionality available for all the algorithms mentioned above.

**Pandas:**

Pandas[21] is a python library for data manipulation and analysis. Pandas has been created to provide a speedy process for data manipulation and analysis. Pandas is exceptionally fast a loading data from csv files. This could be very useful in this project.

**NumPy:**

NumPy[22] is another Python library. NumPy provides functionality for utilising multi-dimensional matrices. Along with utilising multi-dimensional matrices NumPy also provides very high-level mathematical functions associated with matrices. NumPy is extremely fast when dealing with matrices and mathematics.

**Flask:**

Originally the idea for the project was to host the project on a server which provides an array of user-friendly option for the users. This idea was not brought forward to implementation in favour of concentration on the back-end production. Flask[23] was already researched for the use of a front end.

Flask is a python framework used for the creation of web applications. Flask is based off of jinga2 which is a html templating engine. By utilising flask web applications can be easily built and will easily support python back end computation as well as having some functionality for displaying some creations of other python libraries. Graphs from Matplotlib is one example of this.

**Rapid Miner:**

Rapid Miner[24] is very like scikitlearn in that it provides a service allowing machine learning to be undertaken. Unlike scikit it is not a python library but rather it is a standalone application. Installed on the desktop as an application the data to be used is transported into Rapid miner and rapid miner internal functionalities are used to remove outlier, perform machine learning and other data pre-processing techniques among other tasks it can perform.

**The spider technologies**

Python was used to build a spider which included a mechanism for saving the data scraped by the spider. The data was stored in SQLite, but this will be talked about in data storage technologies. The spider used some python libraries which will be talked about here.

**Datetime:**

Datetime is a library in python that allows the use and manipulation of time. This was not originally necessary until testing began when the time was needed to be kept on the spiders run time. Datetime provides functionality that allows the user to set a start and end time then using mathematical equations the programmer deduces the length of time the program was running for.

**Beautiful Soup 4:**

Bs4 as it is more commonly known is used to ascertain information from html. Beautiful soup 4 parses html documents. Essentially speaking bs4 retrieves the html from a website for a developer. Using bs4 the developer can break down the data into smaller sub areas that allows closer examination and manipulation of the html data. It is here that the data required is extracted from the html document.

**Requests:**

Requests is a Python library that provides a service to allow a program to access the internet. The reason for researching requests is because the spider will need internet access to scrape the data required from Daft.ie

**Re:**

The library re is a Python library that provides tools that allow the creation of regular expressions. Regular expressions define a sequence of characters that are then used for searching a string. Re allows the developer to define these characters using re’s syntax. Then re will perform the search for the developer and return a result. This library was researched to utilise the extraction of specifically needed data from the html that beautiful soup 4 parses for the spider.

**Sqlite3:**

Sqlite3 is an API library in Python that allows interaction with an SQLite database. The library allows for the correct syntax to be used in python along with allowing for SQL queries to be utilised. This was necessary as SQLite was where the data was being stored directly after its scraping as will be mentioned later.

Everything mentioned in this section was utilised in the creation of the project’s spider for daft.ie. they were also used when scraping data from Wikipedia.

**Data Storage Technologies:**

With a data set comes the responsibility and requirement of storing the data. The data needs to be stored in a reliable and accessible place. When it was determined that a data set was required it was then ascertained that the dataset would have to be created. This data set was created by scraping data from Daft.ie. Unfortunately, the data set was an urgent necessity and was gleaned at a very early stage of the project with extreme haste placed on its acquirement. For this reason, the spider utilises an implementation of SQLite. The reasons for this will be discussed next.

**SQLite:**

SQLite[25] is a database system written in C but can be used with Python. Using sqlite3 library the data can be stored in the SQLite database. SQLite is a system that is embedded into the final program as opposed to a client-server engine. SQLite was implemented very early in the project without much research done. The reason it was implemented was that the developer had previous knowledge of SQL and there was an urgent need for the spider to be completed.

**Mongo DB:**

Mongo-Db [26]is a database system. It is a document orientated database which has become commonly known as No-SQL. Through its use of documents Mongo DB is considered dynamic. The name of the document format it uses is called B-SON. Mongo DB was specifically built for high performance, scalability and access from a single server deployment to complex multi-site infrastructure. Mongo Db was researched as an alternative to other data storage facilities.

**CSV:**

Comma-separated values. CSV files are delimited files that use commas to sperate values. The data is stored in plain text. Each line of text is a data record and each data record consist of fields. Each of these fields are separated using commas. CSV is a very simple to use system with python. It has its own Python csv library and is portable with R and R-Studio. Saving data in a csv file instead of an SQLite database also saves space. To view a csv file it can be opened in Microsoft excel instead of needing to download a specific program like DB Browser for SQLite.

**JSON:**

JavaScript Object Notation[27] is another file format that data can be stored as. JSON is human readable and uses attribute- value pairs.

**Data pre-processing technologies**

For the majority of pre-processing the techniques used were with the previously mentioned re library but during the pre-processing the data was transferred to csv. Also required was a functionality for acquiring a Latitude and Longitude for each of the properties using their address.

**CSV:**

Csv is a python library that provides functionality to convert python data to csv files. Lists and Dictionaries will be primarily used. Furthermore, csv provides functionality for converting a csv into a python object allowing for the object to be converted to the data type the developer wishes.

**Geopy:**

Geopy provides functionality for geocoding and reverse geocoding addresses in Python.

**Googlemaps:**

Googlemaps[28] is a Python library that utilises the use of Googlemaps to provide the functionality available with Googlemaps.

The above geocoding libraries failed to provide the proper functionality and as such a new way was required. It was decided to access the Googlemaps API through the one provided online by google maps and to use python to pass and receive the addresses and their Latitude and Longitude respectively. To achieve this there was need to access the internet and an internet http link from a Python program.

**Urllib3:**

Urllib3 is a python library that provides access to the internet via the use of URL links.

**Certifi:**

Certifi is a python library that is used to validate trustworthiness of SSL certificates. This library has been extracted from the Requests library mentioned earlier.

**Googlemaps Geocoding API Requests Format:**

Geocoding is the process of converting an address into its Latitude and Longitude values. Google Maps[28] is a mapping service provided by Google that provides maps of the world and some services like geocoding and the distance matrix among other. These can be implemented through JavaScript or other various forms. One such way of geocoding addresses is to pass an address into the following URL link. <https://maps.googleapis.com/maps/api/geocode/outputFormat?parameters>

This format is the Googlemaps Geocoding API Requests Format. The results of sending an address through this link is the addresses Latitude and Longitude position.

**Google Maps API key:**

A google maps API key[28] is needed in order to utilise the google maps platform or any of its linked services. The key is simply a way of tracking the level of usage of a customer and is linked to the customers bank account in order to ensure payment of services rendered. The key comes in a string when provided in the necessary places it provides access to Google Maps. Without a key there can be no geocoding performed on Google Maps.

Unfortunately, for unknown reasons, these attempts of acquiring the Latitude and Longitude for the addresses did not work and another form was needed.

**Geocoder:**

Geocoder is a python library that provides geocoding services. It states in the geocoder documentation that it believes that acquiring geocoding from google maps among others is too difficult due to the format they return their results. Geocoder provides the ability to utilise over twenty different geocoding providers. The provider used in this project was Google Maps. By providing a Google Maps API key the library provide access to google maps in a much simpler and straightforward way to develop.

**Clustering/Classification Algorithm Technologies**

**Math:**

Math is a python library that provides function for mathematical equations and calculations. The use for math will be to utilise linear algebra within the clustering/classification algorithm.

**Operator:**

Operator allows the iteration over python dictionaries using their key as an operator for iteration.

**Data Visualisation and Statistics Technologies**

**Matplotlib:**

Matplotlib[29] is a python library that allows the developer to create graphs. These graphs are created using data provided for by the developer. The graphs in matplotlib can be used to visualise data. Matplotlib comes highly recommended and is widely used in data science. Some graphs that can be created are bar charts, scatter plots, histograms and line charts among others. These graphs can be embedded in certain python applications.

**Statistics:**

Statistics is a python library that provides functionality to determine statistics on data provided by the developer. The developer provides data and statistics uses its inbuilt functions to calculate the desired statistics. Some provided statistics include mean, median and mode.

**R:**

R is a programming language and environment specifically targeting statistical computing and graphics. It is commonly used for creating statistical analysis and data analysis software. R is free to download and use. R can easily handle large data sets and is well supported with a multitude of libraries.

**R-Studio:**

R-Studio[30] was researched as an IDE for developing in the R programming language. It was the only IDE researched as it provided the basic needs to develop in R and no further requirements were necessary. A fantastic attribute of R is the same window display functionality for graphs created in the R language in R-Studio.

**GGPLOT2:**

GG-plot-2 is an R library that provides the capabilities of graphically displaying data in the form of graphs. The research of ggplot2 led to the researcher determining that the graphs in ggplot2 were not only visually more pleasing but also, they were easier to implement and create. There were also some very efficient and accurate utilities for learning ggplot2 provided for this project through the mentoring lecturer.

**Literature Review**

**Real estate price forecasting based on SVM optimised by PSO:**

The objective of the paper by Xibin Wang, Junhoa Wen, Yihao Zhang and Yubiao Wang is to predict the real estate price using ‘a new model based on support vector machine’. They used particle swarm optimisation to optimise their SVM algorithm’s parameters. The paper begins by explaining the regression theory of SVM before delving into an explanation of the model they have based upon PSO-SVM for real estate price forecasting. This is an in-dept explanation where they also provide the following algorithm -‘y=f(x)=∑li=1(− ̨∗i−− ̨i)K(xi,x)+−b’. The paper continues to explain the process that was taken to complete the task and then gives a step by step instruction on how to optimise the parameters of SVM using PSO. A comparative study of the SVM-PSO algorithm against a BP Neural Network algorithm on data from Chingquing shows the SVM-PSO algorithm to possess a higher forecasting accuracy. Finally, the authors state their intention to expand the data set and research ‘how PSO can maintain its good global search and convergence and make the necessary improvements to improve the local search ability , so as to ensure the accuracy of the forecast within the acceptable range.’

**Housing price prediction using neural networks:**

Wang Teng Lim, Lipo Wang, Yaoli Wang and Qing Chang wrote a paper with the intention of forecasting the prices of condominiums in Singapore. A condominium is a ‘building or complex of buildings containing a number of individually owned apartments or houses.’ They apply two models to their problem and then compare the results. The algorithms used are artificial neural networks and auto aggressive integrated moving average. The paper explains the attributes they used to predict the condominium price index and also the condominium asking price. To demonstrate the results of these attempts they showcase them in a table. The results show that ANN holds superiority over the ARIMA model. The authors explain their belief that the input into the algorithms were not to an acceptable standard and in conclusion they state their plan to ‘use more rigorous techniques to select input features’ in the future.

**Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data:**

Byeonghwa Park and Jae Kwon Bae use four separate algorithms to classify the houses in Fairfax County, Virginia. These algorithms are C4.5, RIPPER, Naïve Bayesian and AdaBoost. They then compare these results. To begin they combines some data from different sources together. These sources were multiple listings service, historical mortgage rates and public schools’ ratings. This data was cleaned integrated, extracted, transformed and reduced as explained in the research paper. The four algorithms are then used to classify each house as having a final selling price of higher or lower than the listed price they were shown. The results are demonstrated in tables and show that RIPPER is the superior of the four algorithms. They finish by stating their future work which includes the use of economic variants in the future when predicting house prices.

**Ensemble ANNs-PSO-GA Approach for Day-ahead Stock Exchange Prices Forecasting:**

Fi Xiao, Jin Xiao, Fengbin Lu and Shouyang Wang set out to predict the price of E-exchange Stock one day ahead using artificial neural networks. The introduction of this paper comprises of a detailed discussion of previous applications of artificial neural networks. This discussion presented the authors with an issue. This problem is ‘how to optimize the performance of the base models. They propose to use an ensemble of three different base models, Elman network, generalised regression neural network and wavelet neural network combined by support vector machines neural network in a non-linear way. The paper explains the models in detail. This explanation is very extensive and includes several algorithms. They then explain how they optimise the base models with another detailed and expansive explanation. The final parts of the paper begin to explain the process of the prediction, beginning with data preparation and pre-processing. They use a time dependency included in the data and must calculate the lag order of the time series before continuing to simulation and predicting. As expected the results of simulation and predicting receive another exhaustive, extensive and detailed discussion with the aid of some diagrams and tables. Overall the results show that the prediction prowess of the non-linear ensemble model from the research has a better annual return rate than the algorithms used as comparison.

**Decisions Made**

In this section the decisions made on the areas researched for the projects structure along with the technologies decided to be used will be discussed. The decisions will be stated and then a reason for the decision will be provided with enough relevant detail without exhausting the topic provided. Any initial decisions that were made but overturned will also be detailed here.

**Project Structure Decisions**

By utilising the areas researched a project structure and plan was assembled and this structure follows the headings provided which will all be discussed.

**What we will predict:**

Having researched the topic or area of property price prediction it was ascertained that this meant predicting the sale price of a property based upon the attributes available for that property. For this reason, it was decided to predict the sale price of the house based off of its attributes.

To supplement this prediction, it was decided that the project would aim to predict the future price of a property by utilising historical data.

**Where we will get the data and how:**

To predict the price of a property some data was needed and the best way to retrieve data was decided to be Daft.ie. Daft.ie provides enough data to glean attributes and a price for the property. However, the price provided by Daft.ie was not considered to be of complete use for the prediction process. It was decided instead that the actual selling price would be required. This selling price was available from the property price register.

The property price register can also provide a date of sale. Combining the data from both the property price register and daft.ie would be sufficient for both prediction goals. This resulted in the decision to acquire the data from both locations and combine it. A further requirement for data was Latitude and Longitude locations for each property. Googlemaps was chosen for this purpose as will discussed later.

**How the project will utilise data science:**

It was decided to utilise almost all of the data science disciplines mentioned above. Mostly these decisions were made out of necessity to complete the project.

**The machine learning algorithm the project will use:**

Q-Learning was chosen as the machine learning algorithm to be used in the project. Q-learning would provide an innovative approach to the problem, it would provide an interesting venture for the developer and is relatively simple with regards to the use of formulae. These points are insignificant points for the decision to use q-learning but did have some small bearing on the final decision.

Ultimately the decision to use q-learning was based off the fact that the data made available through the project’s endeavours did not offer up a complete data model. Q-learning as discussed earlier is a model-free algorithm. This means that the q-learning algorithm does not need to see a complete data model and would thrive under the circumstances provided by the data that had be retrieved. There were no further model free algorithms researched because q-learning was so ideal for the situation at hand.

Initially there was a different idea in place to combat the problem of predicting property price. This idea was put in place before the data was identified as not having a complete model of the environment.

This idea involved the use of a random forest to classify the areas of a county into categories based upon the relative price of property in the area and to then use neural networks to categorise the properties based off of their attributes. Then using the area and the newly acquired attribute category a linear regression model would be used to predict the properties price. This theory as mentioned was scraped in favour of q-learning.

**How the data will be retrieved:**

The decision was made to retrieve the daft.ie data by using a spider. The spider would first attain the links to each individual property and would then feed these links to a scraper to retrieve the levant data and save it to a database.

The data from the property price register is available for download into a csv file directly from the website.

**How the areas will be split into sub areas:**

The areas effect on a properties price was far too extreme to ignore. This called for a clustering algorithm. Unfortunately, the algorithms which were identified as being the best fit were researched and ultimately deemed useless for the purpose needed in this project. These algorithms were density-based spatial clustering of applications with noise and k-means clustering. The properties needed to be clustered based off of their Latitude and Longitude location, but it was decided that this was not enough. The algorithm needed to also consider the price of the property when clustering the properties. The approach of clustering on value and location separately was analysed but the decision to develop a custom-built clustering algorithm was made.

**Formulae needed from graph theory:**

All the formulae mentioned above were decided to be used in the custom-built clustering algorithm.

**Technology Decisions**

**General use technologies:**

Tech to be used in the project:

The general tech that will be used in this project are as follows:

* Windows
* Python
* PyCharm
* Python Interpreter and pip

Windows was chosen out of necessity as mentioned earlier. Python was chosen because between itself and R the developer had the most knowledge of Python so it made absolute logical sense to use Python. As PyCharm was the IDE the developer was familiar with and with no underlying issues identified with PyCharm it was chosen as the IDE for Python. As mentioned Anaconda was the original package manager, installer and interpreter chosen. This was replaced with the Python interpreter and pip when it started to cause issues.

Git and GitHub were chosen as the version control software for the project. In order to visualise the SQLite data a browser called DB Browser for SQLite was chosen as it was the first software identified as being user-friendly that satisfied the requirements. This was a straightforward decision process as it was not critical to the project.

Tech that will not be used:

Anaconda as mentioned was decided not to be used. Scikitlearn was rejected as a solution to utilising machine learning, as was Rapid Miner. These technologies were rejected as the decision to write a machine learning algorithm from scratch rendered these technologies of no use to the project. Flask was removed from the project when the decision to concentrate on the backend was made. Finally, there was intent to use Pandas and NumPy but as the project progressed there was no use found or required and they ended up being left out of the project.

**Spider technologies:**

The decisions made for the spider were mentioned above as all the tech mentioned were actually utilised in the spider. There was no use for any further technologies when developing the spider. The technologies used were:

* Datetime
* Beautiful Soup 4
* Requests
* Re
* SQLite3

**Data storage technologies:**

It was decided to use SQLite and CSV in the project for storing the data and transferring it from file to file or from python to R.

The decision to use SQLite was made because the developer was forced to decide on a technology in a hasty time frame and with some previous working knowledge of SQL it made sense to use a SQL database. SQLite was chosen as the SQLite database due to its portability and ability to be used without a server or server technology. Using an SQL database made the decision to scrap Mongo DB an easy one.

As previously mentioned the data that was downloaded from the property price register was in CSV format and as such it made no sense to transfer this data to JSON. CSV was chosen as the lightweight data storage format.

**Data pre-processing:**

The data that was retrieved from both sources needed a lot of pre-processing in order to be ready for analysis. The following technologies were chosen to be used in this process:

* CSV
* Geocoder
* Google Maps API Key

These technologies were the best fit for the task at hand as mentioned above in their description. The remaining technologies mentioned in this area were not used as mentioned above. There were many other technologies used in this section that were already mentioned in previous sections.

**Clustering algorithm:**

Math and operator were both chosen to be utilised in the custom-made clustering algorithm.

**Data visualisation and statistics:**

The resources made available by the mentoring lecturer made the decision to use GGPLOT2 a straightforward decision. By making this decision R was made necessary as well as an IDE which was R-Studio. Matplotlib and statistics were removed from the project.

**Bibliography**

[1] House Prices: Advanced Regression Techniques | Kaggle. 2019. House Prices: Advanced Regression Techniques | Kaggle. [ONLINE] Available at: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>. [Accessed 01 November 2018].

[2] The Balance. 2019. Making Money From Buying Stocks. [ONLINE] Available at: <https://www.thebalance.com/how-to-make-money-from-buying-stocks-357330>. [Accessed 01 May 2019]. [3] Karen Hao. 2019. What is machine learning? We drew you another flowchart - MIT Technology Review. [ONLINE] Available at: <https://www.technologyreview.com/s/612437/what-is-machine-learning-we-drew-you-another-flowchart/>. [Accessed 01 November 2018].

[SVM] Rohith Gandhi. 2019. Support Vector Machine — Introduction to Machine Learning Algorithms. [ONLINE] Available at: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>. [Accessed 01 November 2018].

[5] Python.org. 2019. Welcome to Python.org. [ONLINE] Available at: <https://www.python.org/>. [Accessed 01 November 2018].

[6] JetBrains. 2019. PyCharm: the Python IDE for Professional Developers by JetBrains. [ONLINE] Available at: <https://www.jetbrains.com/pycharm/>. [Accessed 01 November 2018].

[7]Kaggle: Your Home for Data Science. 2019. Kaggle: Your Home for Data Science. [ONLINE] Available at: <https://www.kaggle.com/>. [Accessed 01 November 2018].

[8]Search Ireland's No. 1 Property Website | Daft.ie. 2019. Search Ireland's No. 1 Property Website | Daft.ie. [ONLINE] Available at: <https://www.daft.ie/>. [Accessed 01 November 2018].

[9]Gruss, J., 2016. data science from scratch. 3rd ed. Sebastopol, United States: O'Reilly Media, Inc, USA.

[10About logistic regression | IBM. 2019. About logistic regression | IBM. [ONLINE] Available at: <https://www.ibm.com/analytics/learn/logistic-regression?S_PKG=-&cm_mmc=Search_Google-_-Hybrid+Cloud_Business+Analytics-_-WW_IE-_-logistic+regression_Exact_-&cm_mmca1=000000OA&cm_mmca2=10000380&cm_mmca7=1007846&cm_mmca8=kwd-300037924682&cm_mmca9=_k_Cj0KCQjwh6XmBRDRARIsAKNInDG5Mz5FwCD3o5QJuL4K13OUz3v4p8fUDYCgr1Z76bTk56pIzKlU7OUaAhxdEALw_wcB_k_&cm_mmca10=335197109958&cm_mmca11=e&mkwid=_k_Cj0KCQjwh6XmBRDRARIsAKNInDG5Mz5FwCD3o5QJuL4K13OUz3v4p8fUDYCgr1Z76bTk56pIzKlU7OUaAhxdEALw_wcB_k_|479|404725&cvosrc=ppc.google.logistic%20regression&cvo_campaign=000000OA&cvo_crid=335197109958&Matchtype=e&gclid=Cj0KCQjwh6XmBRDRARIsAKNInDG5Mz5FwCD3o5QJuL4K13OUz3v4p8fUDYCgr1Z76bTk56pIzKlU7OUaAhxdEALw_wcB>. [Accessed 01 November 2018].

[11]Analytics Vidhya. 2019. Understanding Support Vector Machine algorithm from examples (along with code). [ONLINE] Available at: <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>. [Accessed 01 November 2018].

[12]Random forests - classification description. 2019. Random forests - classification description. [ONLINE] Available at: https://www.stat.berkeley.edu/~breiman/RandomForests/cc\_home.htm. [Accessed 01 November 2018].

[13]Neural Networks. 2019. Neural Networks. [ONLINE] Available at: <https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html>. [Accessed 01 November 2018].

[14] ADL. 2019. An introduction to Q-Learning: reinforcement learning. [ONLINE] Available at: <https://medium.freecodecamp.org/an-introduction-to-q-learning-reinforcement-learning-14ac0b4493cc>. [Accessed 01 November 2018].

[15] Reinforcement Learning – Q Learning. Available at: https://www.youtube.com/watch?v=1XRahNzA5bE (Accessed: 18 November 2018).

[16]Anaconda. 2019. Home - Anaconda. [ONLINE] Available at: https://www.anaconda.com/. [Accessed 01 November 2018].

[17]GitHub. 2019. The world’s leading software development platform · GitHub. [ONLINE] Available at: https://github.com/. [Accessed 01 November 2018].

[18]Git. 2019. Git. [ONLINE] Available at: https://git-scm.com/. [Accessed 01 November 2018].

[19]DB Browser for SQLite. 2019. DB Browser for SQLite. [ONLINE] Available at: https://sqlitebrowser.org/. [Accessed 01 November 2018].

[20]scikit-learn: machine learning in Python — scikit-learn 0.20.3 documentation. 2019. scikit-learn: machine learning in Python — scikit-learn 0.20.3 documentation. [ONLINE] Available at: https://scikit-learn.org/stable/. [Accessed 01 November 2018].

[21]Python Data Analysis Library — pandas: Python Data Analysis Library. 2019. Python Data Analysis Library — pandas: Python Data Analysis Library. [ONLINE] Available at: https://pandas.pydata.org/. [Accessed 01 November 2018].

[22]NumPy — NumPy. 2019. NumPy — NumPy. [ONLINE] Available at: https://www.numpy.org/. [Accessed 01 November 2018].

[23]Welcome | Flask (A Python Microframework). 2019. Welcome | Flask (A Python Microframework). [ONLINE] Available at: http://flask.pocoo.org/. [Accessed 01 November 2018].

[24]RapidMiner. 2019. Lightning Fast Data Science Platform for Teams | RapidMiner©. [ONLINE] Available at: <https://rapidminer.com/?utm_source=Google+Ad&utm_medium=cpc&utm_content=brand+>. [Accessed 01 November 2018].

[25]SQLite Home Page. 2019. SQLite Home Page. [ONLINE] Available at: https://www.sqlite.org/index.html. [Accessed 01 November 2018].

[26]MongoDB. 2019. The most popular database for modern apps | MongoDB. [ONLINE] Available at: <https://www.mongodb.com/>. [Accessed 01 November 2018].

[27]JSON | Couchbase. 2019. JSON | Couchbase. [ONLINE] Available at: https://info.couchbase.com/couchbase\_json.html?utm\_source=google&utm\_medium=search&utm\_campaign={campaign}&utm\_keyword=json&gclid=Cj0KCQjwh6XmBRDRARIsAKNInDEAuWOmN-mkto9nw7c-6I9JbzKfiTkFA5aBjyEPU577Ou9mdQRbnRkaArloEALw\_wcB. [Accessed 01 November 2018].

[28]Google Cloud. 2019. Geo-location APIs  |  Google Maps Platform  |  Google Cloud. [ONLINE] Available at: <https://cloud.google.com/maps-platform/?utm_source=google&utm_medium=cpc&utm_campaign=FY18-Q2-global-demandgen-paidsearchonnetworkhouseads-cs-maps_contactsal_saf&utm_content=text-ad-none-none-DEV_c-CRE_277743457148-ADGP_Hybrid+%7C+AW+SEM+%7C+SKWS+~+Mapping+APIs-KWID_43700037731655637-kwd-297933066873-userloc_1007846&utm_term=KW_%2Bmaps%20%2Bapi-ST_%2Bmaps+%2Bapi&gclid=CJXHgNa8-uECFURqGwodJ0cOiw>. [Accessed 01 November 2018].

[29]Matplotlib: Python plotting — Matplotlib 3.0.3 documentation. 2019. Matplotlib: Python plotting — Matplotlib 3.0.3 documentation. [ONLINE] Available at: <https://matplotlib.org/>. [Accessed 01 November 2018].

[30]RStudio. 2019. Open source and enterprise-ready professional software for data science - RStudio. [ONLINE] Available at: https://www.rstudio.com/. [Accessed 01 November 2018].