

Property Price Prediction

Alan Cahill

C00174533

Lecturer:

Dr. Lei Shi

Functional Specification

**Table of Contents**

**Abstract3**

**Project Functionality4**

Scrape From Daft.ie**4**

Classification**4**

Q-Learning**5**

**Target Groups6**

**Metrics7**

**Timeline9**

**Abstract**

Property Price Prediction is a program that is built in Python with the intention of predicting the prices of properties. The predicting of prices of a property will never be an area that could achieve a 100% accuracy hence the statement that the project intends to predict prices. The properties prices are ‘predicted’ using historical data that was gleaned from online. A Q-Learning algorithm is used to predict the values after the data has been properly pre-processed.

The functional specification will detail the project in more depth whilst also providing further analysis on the following topics:

* Project functionality
* Target Groups
* Timeline
* Metrics

**Project functionality**

Property Price Prediction as previously stated is a program written in Python with the intention of predicting the selling price of properties in Ireland. More specifically the project provides a proof of concept by applying its functionality to two areas in Ireland:

* Dublin 15
* Cork City Centre

The functionality of the project falls under the following headings:

* Scrape Data From Daft.ie
* Pre-Processing
* Property Classification
* Q-Learning

**Scrape from Daft.ie:**

The program begins by running a spider. The spider is designed to scrape information from daft.ie. To begin the spider accesses the properties for sale page. This is the first page of numerous pages that contains all the properties registered with daft that are for sale in Ireland. This amounts to close to 30,000 properties. With each property comes a link to the properties page which the spider scrapes and appends to a list. This list is then traversed, and each link is used to scrape the details of the property of the link. The data is saved and added to a database.

Pre-Processing:

The data retrieved from daft is not at an acceptable level for use or even readability. The program uses pre-processing techniques to ensure the data becomes acceptable. The areas of pre-processing utilised are as follows:

* Cleaning
* Transformation
* Integration

The data required substantial cleaning and the program obliges. Without getting into specific details, after the cleaning the data provides no problems with a clean and polished look.

The data extracted only provides for some of the necessary information needed in the project. The program transforms the data by utilising existing data to create and determine new information and data.

The project integrates the daft data with the data set gleaned from the property price register.

**Classification:**

The data required some further processing. This processing amounted to classifying the properties into sub areas that would provide further insight into the value of the property. To achieve this an algorithm was designed and utilised. The algorithm follows the following steps to achieve classification:

1. Calculate a centre node
2. Plot an x and a y axis from the centre node
3. Assign all properties their relevant compass position relative to the centre node
4. Use a mathematical algorithm to find the border point of each compass position
5. Assign every property between the centre node and the border point as being in that particular classifier
6. Repeat steps 1 to 6 until 90% of the properties have been assigned

**Q-Learning:**

Finally, the project splits the data 80%-20% and passes 80% data into a q-learning algorithm. This trains the algorithm to predict the price of a particular property. The program then uses the 20% data minus the price of each property to predict the price of the properties in question through the use of the q learning algorithm.

**Target Groups**

The target group for this project is property auctioneers. Normally an auctioneer would appraise a property that has been presented by a customer. This project would hope to remove this task from the auctioneer and allow them to concentrate on the sale of the property. This could be a huge benefit to some of the bigger and more widespread auctioneers but unfortunately the smaller auctioneers would most likely view this project as a direct competitor. The reasons for this is the bigger auctioneers would save huge amounts of time and manpower by utilising the algorithm and would not loose too much revenue from the changeover. The reason being the main income for auctioneers like these would be commission on sales. The smaller auctioneers would unfortunately be more dependant on the income from appraisals which could possibly cover running costs among other aspects.

In conclusion the bigger and more widespread auctioneers are the target groups with the smaller auctioneers not specifically targeted but will always be welcome to utilise the project.

**Metrics**

The project’s success is based on one major issue, how accurate are the predictions made by the project. Providing accurate predictions is the fundamental plan of the project and the success of the project should be calculated with this goal in mind. However, calculating the projects success should not ignore the other work the project has undertaken. For this reason the projects success must include the following:

* Was the data retrieved from daft.ie?
* Is the classifier accurate?
* Is the Q-Learning algorithm accurate?
* How much work was achieved in the pre-processing?
* How was the functionality undertaken in the project?

Was the data retrieved from daft.ie:

A success scenario in this instance is the data was retrieved from daft.ie. The data must be complete enough for the project to continue from this point.

Is the classifier accurate:

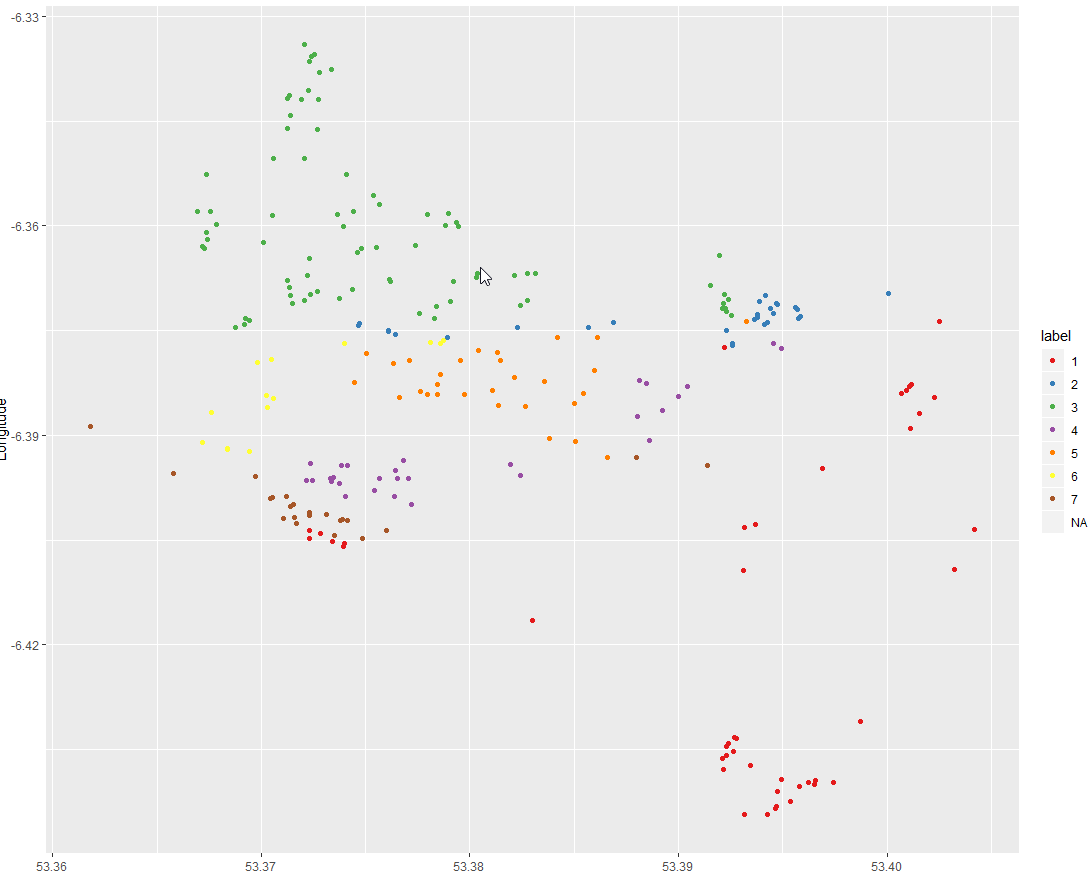
This section provides an issue. How can the accuracy be tested?! There exists only one solution to this problem. There exists a graph as show in Fig.1. This graph can be used to determine the accuracy of the algorithm. To determine the accuracy would require human intuition. A diagram with erratic positioning of nodes of similar colour or with nodes of similar colour overlapping with each other. 

Fig.1

Is the Q-Learning accurate:

The accuracy of the q-learning predictions can determine the success of the project. These can be tested to determine their accuracy. The project used the coefficient of determination but this proved to not be useful as it is a formula designed for use with a regression line. Other ways could be utilised that the project did not have time to do. These include assessing the accuracy of each individual prediction and finding an average of the total. To utilise this procedure all outlier would need to be removed. This is because some of the properties predicted were properties that were of extremely low quality and as such would provide obscure results.

**Timeline**

The timeline of the project was October 1st to April 10th. The author however, begun the research into data science around the 1st of August before college had begun. This means in total the authors timeline is August 2018 – April 2019. The timeline below is how the project is planned to be carried out.

October – Research on Data Science

October 28th = November 3rd – Develop the Crawler/Spider.

November – Continue Research

November/December – Pre-process the data

December- Research Q-Learning and learn to implement

January – Analyse the data using stats and graphs

February – Complete uncompleted documents

March – Begin Implementing the Q-Learning Algorithm

April – Finish the Q-learning algorithm

April – Complete a user friendly aspect to utilise the Project.

This is the timeline set out to complete the project. This timeline is a very high level design for implementing a project. There are too many sub areas to mention and the author did not specify any specific dates for longer term goals, instead leaving a flexibility in place throughout the projects life cycle. The smaller more specific goals that can be achieved in a smaller time frame will acquire some deadlines the author deems appropriate throughout the project and its completion.