EPSRC
Engineering and Physical Sciences
Research Council

InFoMM
Industrially Focused
Mathematical Modelling

EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling

Application of Deep Learning Techniques to Commercial Real Estate Appraisal

Giuseppe Ughi

UNIVERSITY OF
OXFORD

NewRock
CAPITAL MANAGEMENT

Contents

1 Introduction **1**

Problem Definition 1

Glossary of terms 2

2 Regression Techniques **2**

3 Deep Learning **3**

4 Data **4**

5 Results **5**

6 Discussion, conclusions & recommendations **6**

1. Introduction

New Rock Capital Management (NRCM) is a firm focused on real estate analytics, portfolio analysis, and risk management. One of its interests lies in the use of publically available data for the appraisal of these kind of Commercial Real Estates. Currently, the appraisal method for commercial estates is problematic for several reasons: the assets are illiquid and traded infrequently, and there is no centralised market for the price discovery. For this reason, the appraisal is usually assigned to some agent who will have to study the location in detail and understand its features. This usually takes a lot of time and, besides being manually intensive and expensive, the valuation is subjective to the agent.

The availability of data and recent advances in computing power could lead to a computerised valuation system in which the valuations become more objective, fast, and accurate. The large amount of data suggests using machine learning algorithms, which have achieved better accuracy than trained humans in recognition tasks in recent years. These algorithms have, as an input, a very big dataset from which they are able to extract the main patterns that define some categories. When new data are introduced, these algorithms are then able to categorise them.

We want to automatise the appraisal of commercial real estates in a similar way to residential real estates.

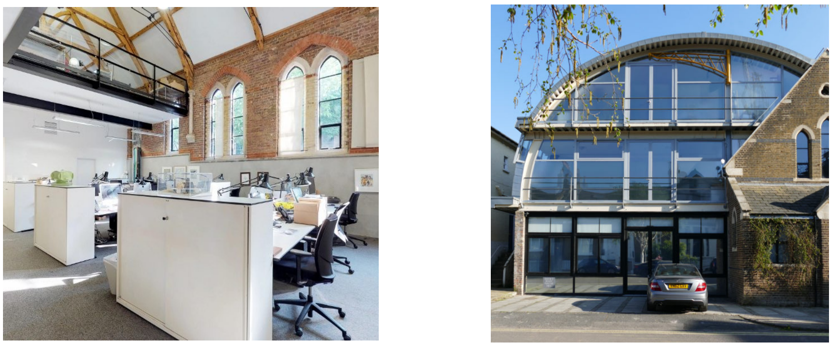
Concerning residential real estates, the automatised valuation of properties is already commonly used. Companies such as Zillow offer to their clients an evaluation based on the neighbourhood and the features of the house such as the number of rooms and toilets. These automatised appraisals are now often used as a guidline to buyers since these estimates in average have an error between 5 − 10%. Given the more complicate nature of commercial real estates, no valuation is offered online.

Our aim is to explore how to implement an appraisal algorithm for commercial real estates using machine learning techniques. In particular, we aim to apply image analysis algorithms to replicate the subjectivity involved in appraisals.

Problem Definition

One of the main issues with which we have to deal during a valuation is the subjectivity of how attractive a property is for a buyer. For example, if we were to open a start up company in a technological environment, we would value the most open spaces that allow the people to exchange ideas easily and without barriers, and thus would pay an extra for this feature. On the other hand, if we were to open a GP surgery, we would be searching for closed rooms in order to ensure patient confidentiality.

To tackle the appraisal subjectivity, we will analyse property images, such as the ones shown in Figure 1, with the help of deep learning. This is a class of algorithms that allows us to identify which category an image belongs to. In classification tasks, deep learning achieves extremely high accuracy, sometimes better than trained humans. These algorithms are also often able to highlight the principal features that define the classes.



(a)

(b)

Figure 1 – Examples of the images that we have access to for each property.

Glossary of terms

- **Absolute Error:** This is the absolute value of the difference between predicted and real values.
- **Relative Error:** This is how much a prediction is wrong relatively to its dimension. Given the prediction value of an estate, p, and the real value, r, the relative error, e, is computed using
$$e = \frac{r - p}{p}. \tag{1}$$
This implies that the real value of an estate is $r = (1 + e)p$.
- **Heteroschedasticity:** This is property of a dataset in which sub-populations of the set have different variabilities from others. Here "variability" describes how the observed values of a population are dispersed from the expected one.
- **Linear Regression:** This is a way to explain the relationship between a dependent variable and one or more independent variables using a straight line, see Figure 2.
- **Support Vector Regression:** This is a more sophisticated method than the linear regression as it tries to explain the relationship between a dependent variable and multiple independent variables. It considers bell shaped functions centered in the sample points, as shown in the example in 2 dimensions in Figure 5.
- **Mean square error:** This is the average of the squares of the errors.

2. Regression Techniques

We use regression techiniques to relate the value of an estate to different independent variables.

Regression models define a relation between dependent and independent variable. The parameters that define this relation are computed by minimising on a sample set the difference between the predicted and real dependent variables. For example, in Figure 2 we consider the linear regression in one dimension; the slope and the height of the dotted green line are the parameters of this kind of regression. These parameters are obtained by minimising the mean square error.

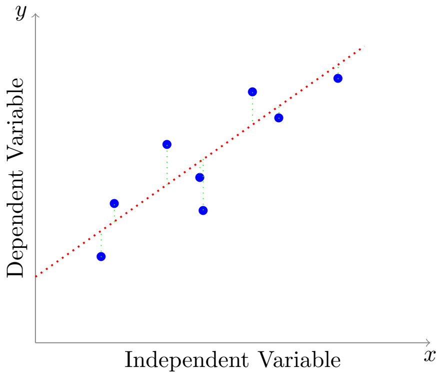


Figure 2 – An example of a linear regression (red dotted line) for a set of data (blue points) with the reppresentation of the error for each valuation (vertical dotted line).

When considering the data relevant to our problem, shown in Figure 3, we observe heteroschedasticity from the relation between the rent of a real estate and its actual total area. To improve the understanding of the relation it is necessary to consider a logarithmic scale; as we see in Figure 4, a model can be implemented more easily in this scale.

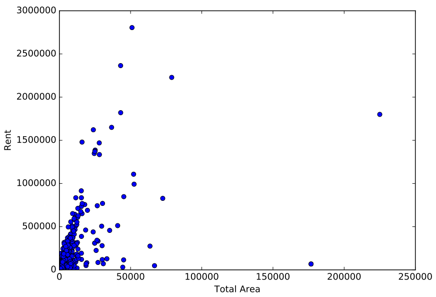


Figure 3 – Representation of the real rent of an office in London over the predicted rent on a linear scale.

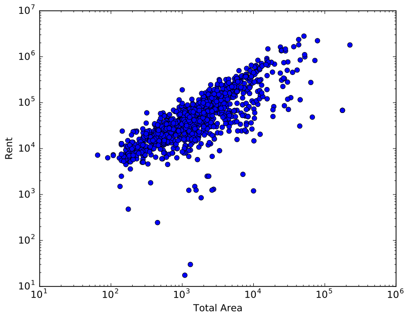


Figure 4 – Representation of the real rent of an office in London over the predicted rent on a logarithmic scale.

However, considering the logarithmic scale is not enough to improve the error since the linear regression does not describe the relationship between the different variables well enough. To allow multiple inputs that could interact in a complex way, we consider *support vector regression*. In this regression, we consider bell-shaped functions, centered at the sample point, with width chosen such that prediction error of the regression is minimised. It is important to note that, in this kind of regression, the error in the sample only influences the width of the bells when it exceeds a prescribed parameter ε. In Figure 5, we show an example in two dimensions of the model generated with this method given two sample points; on the vertical and lateral plane we plot the projection of the maximum values obtained for each bell.

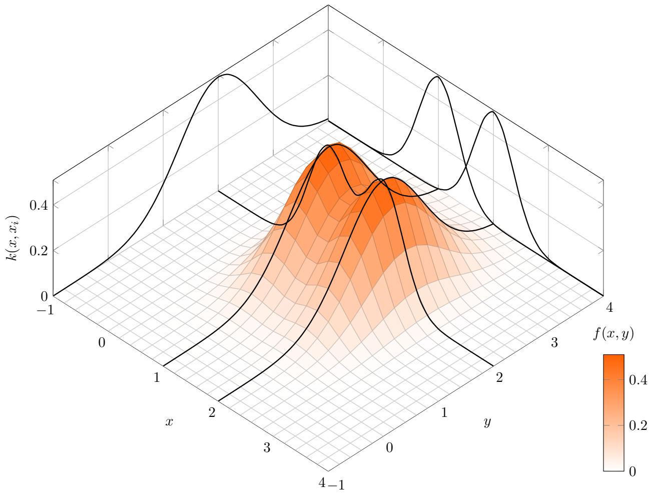


Figure 5 – An example of Support Vector Regression in two dimensions with two sample points located at (1,2) and (2,2). On the vertical and lateral plane we project the shape of the bells centered in the samples.

3. Deep Learning

State of the art deep learning algorithms achieve superior accuracy than humans.

As mentioned in the Introduction, deep learning is a class of algorithms that allows us to assign images into predefined categories. Deep learning has achieved extremely high accuracy in categorising objects, and in certain cases its accuracy is even superior to trained humans.

The chief characterisitc of these algorithms is that they are hierarchical, that is that, iteratively, an input is processed and more abstract features are retrieved. For example, in Figure 6 we present a possible architecture for a deep learning algorithm, also called artificial neural net (ANN), and we see that, in the first iteration, or layer, the image is used as an input and information regarding body parts, as the eye, are obtained. In the following layers, these parts are studied in more detail and new features are extracted.

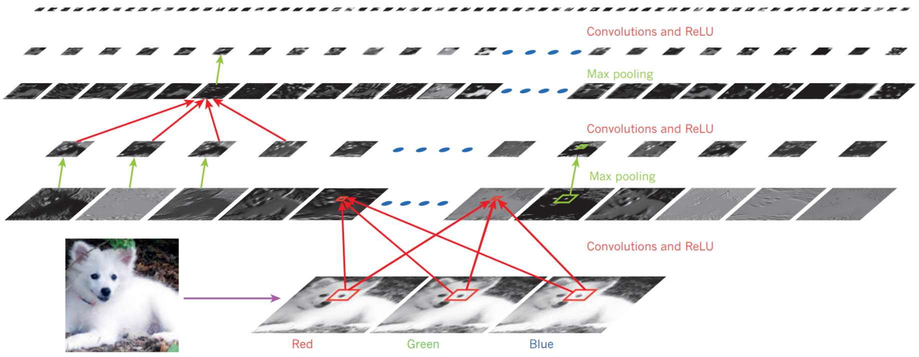


Figure 6 – Presentation of the architecture of a simple ANN. The image of a fox is given as an input and as the layers increase the more abstract features are found.

The proper classification of an image is computed in the last layer which is defined by a list of numbers. This list has as many elements as the number categories that we want the algorithm to identify. Each element of the list is related to one of the classes and the element of the list which is heighest in value corresponds to the most likely category. For example, if the n-th element of the list has the maximum value, then the n-th category is the most likely to be the real one.

A very important consequence for the classification of images following this strategy is that any image will be classified in only one of these classes. For example, if we want the ANN to distinguish different kind of flowers and we give as an input the image of a car, the algorithm will say that the car is a specific flower, and there is no way to find this error other than by checking in person the results and the input images. Furthermore, the quality of the regression is strongly related to the way we train the algorithm. As in the regression we need a set of data to find the optimal parameters. However, in deep learning we also need the inital set of images to be assigned to their class in order to find the value of the parameters that minimises the classification error. As a consequence, gathering a good sample of training images is central to a good accuracy.

4. Data

We focus on the valuation of offices in London and we distinguish the information that we use from the following sources

- Brochures: The brochures deliver three main pieces of information: the asking price for the estate; its position, dimension, type of use among other information; and often its images.
- Borough Information: The city of London provides, for every borough, the population density, the average age, the employment rate, the household median income, the number of works, the job density, the crime per thousand, and the percentage of green space.
- Means of Transport: We determine the nearest underground station to an estate. Given the structure of the underground, we also compute the "betweenness" centrality of each station, which is a measure of how near a station is to the whole system, see Figure 7.

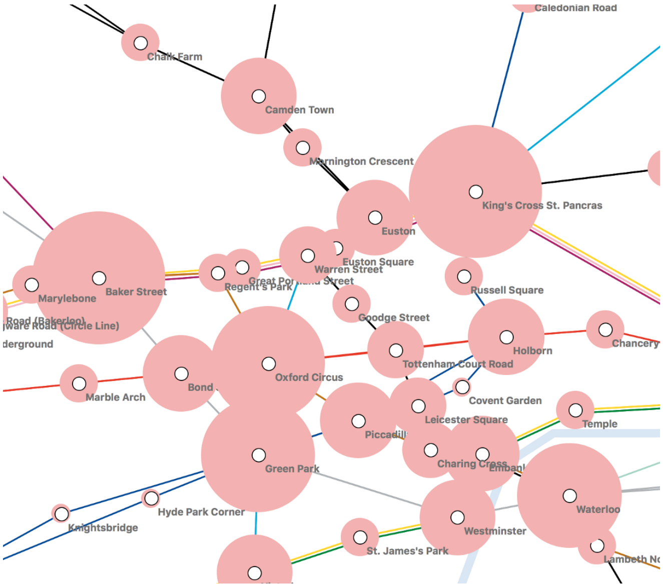


Figure 7 – Network representation of the London underground. The size of the nodes which correspond to the station is proportional to how near the station is to the rest of the system.

- Residential Price: From the different websites used to sell residential real estates, we obtain the price per square foot in each borough in London.
- Rateable Value: The Valuation Office Agency (VOA) is a government body in England and Wales that is in charge of estimating the value of commercial properties with the purpose of the Council Tax formulation. The Rateable value for every single commercial real estate is gathered in a database which is publicly available.

When computing a regression, we need to have independent variables. As a consequence, all the images of properties that we have access to will be analysed with deep learning algorithms to obtain a quality measure as a further input to the regression. We process and generate the data relating to several thousand of properties.

5. Results

We first compute a support vector regression of the value of the estates over all the data not related to images. With this techinique we are able to observe a strong relation between the actual rent of a commercial real estate and the forcasted one from our regression.

When considering the images, we implement deep learning algorithms for different classifications, using images obtained from the brochures. In these documents, the images are not always inherent to the building since, for example, sometimes restaurants in the neighbourhood are shown. Thus, our first step is to remove irrelevant images. We then focus on understanding the quality of the interiors, expecially of the work-related spaces, and we discard images of toilets or hallways. Once this last classification is complete, we focus on giving each working environment a quality measure between one and three according to how functional the space is for the exchange of ideas between colleagues.

We relatively improve by 15% the accuracy by considering qualitative measures of the images.

By considering the image based data, we are able to improve the median relative error of our forecast on the rent of commercial estates by 15% . The final distribution of the relative error of the rent prediction is displayed in Figure 8. We see that the error is centered in 0, but unfortunately there are few cases in which the price is wrong by more than 50%.

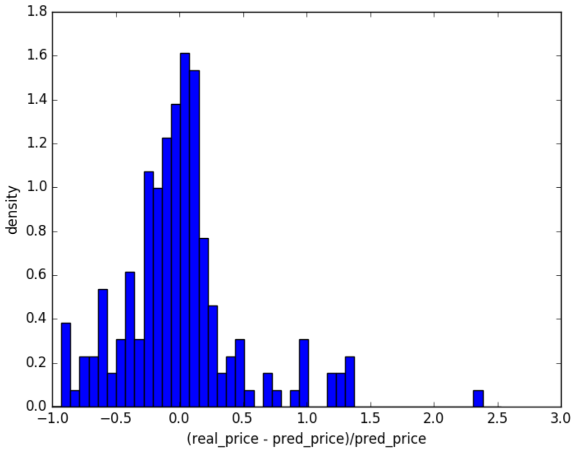


Figure 8 – Relative error distribution for the testing set. The median error is 0.047 and the standard deviation is 0.5621

6. Discussion, conclusions & recommendations

We have explored the possibility of computerising the appraisal of commercial real estates. For this purpose, we first generated a model built on data available on the internet. Afterwards, we used deep learning algorithms to find relevant images and assign a quality measure to the estates.

We showed that machine learning algorithms can be implemented to aid the appraisal of real estates. Moreover, our results are promising since they show that, by using images, it is possible to improve the quality of the regression as these data disclose some of the subjectiveness involved in the appraisal of a real estate. The contribution of the image analysis should in the future further improve the regression accuracy as we will consider all the images in the brochures, not the working enviornments only, and, thus, we should get a quality measure of the area where an office is located.

Richard Hodgson, CEO of New Rock Capital Management, commented "*Giuseppe's work has been vital to the development of a new project idea. His work helped identify the issues, collect relevant data, build analytic tools and explore the use of deep learning for image classification. The resultant understanding demonstrates the feasibility of such methods and provides a focus for exploration in a much longer research project which we will jointly pursue*".