www.mecsj.com

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)

Issues (62) 2023

ISSN: 2616-9185

HOUSE PRICE PREDICTION USING DEEP LEARNING AND COMPUTER VISION TECHNIQUES

ROLA M. ELBAKLY[1] , MAGDA M. MADBOULY,[2] SHAWKAT K. GUIRGUIS[3]

[1]Msc. Student, Institute of Graduate Studies and Research, Department of IT, Alexandria, Egypt

[2]Lecturer, Institute of Graduate Studies and Research, Department of IT, Alexandria, Egypt

[3]Professor, Institute of Graduate Studies and Research, Department of IT, Alexandria, Egypt

E-mail: [1]rola.elbakly@gmail.com, [2]mmadbouly@hotmail.com, [3]shawkat_g@yahoo.com

الملخص

لطالما كان التنبؤ بالأسعار موضوعًا مهمًا للأفراد والشركات. في الاقتصاد العالمي اليوم ، أصبح سوق الإنترنت هو المعيار الجديد ، مما زاد من الحاجة إلى تطبيقات جديدة دقيقة وسهلة الاستخدام للتنبؤ بالأسعار. الهدف في دراستنا هو فحص أهمية صورة المنتج في التنبؤ بسعره. نتوقع أولاً

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
Issues (62) 2023
ISSN: 2616-9185
www.mecsj.com

أسعار العقارات باستخدام بيانات الصورة فقط عن طريق الTransfer Learning. وثانيًا ، نصمم نموذجًا هجينًا جديدًا يستغل السمات العددية وصورة المنزل من أجل التنبؤ بالسعر بدقة أعلى. نجح نموذجنا الهجين في تقدير قيمة المنازل بدقه تصل الى 0.91 .

الكلمات المفتاحية:الشبكات العصبية العميقة ، الانحدار ، التنبؤ بالأسعار ، الرؤية الحاسوبية ، الأنظمة الهجينة

ABSTRACT

Price prediction has always been a crucial topic for individuals and businesses. In today’s new global economy, the internet market has become the new norm, raising the need for new accurate user-friendly price prediction applications. In our study the aim is to examine the significance of product image in its price prediction. We first predict real estate prices using image data only utilizing transfer learning. And second we design a new hybrid model that exploits both numerical and image attributes of a house listing in order to predict the price with higher accuracy. Our hybrid model successfully estimated the value of houses with a R2 of 0.91.

Keywords: *Deep Neural Networks, Regression, Price Prediction, Computer Vision, Hybrid Systems*

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
Issues (62) 2023
ISSN: 2616-9185
www.mecsj.com

1. INTRODUCTION

Any product on sale is presented on a website using: a textual description, a product specifications table and several images, showing the purchasable item from different angles and views. In this study we aim to contribute to the growing area of computer vision research by exploring the importance of image data on the price prediction accuracy. This paper adds to the current literature by offering a hybrid deep learning system that employs the given numerical and image data to predict the price of real estate listing with high accuracy. The objective of this paper is to further investigate the significance of utilizing image data on price prediction accuracy. By far, the currently most popular methods of price estimation used in research are the tree-based approaches. This is due to the high accuracy and stability achieved using these approaches, the downside though, is that its use is limited on numerical and categorical data. Our goal is to use Random Forest (RF) on numerical and categorical data together with Transfer Learning to design a new hybrid model that can deliver highly accurate price estimations. The overall structure of the study takes the form of six sections, including this introductory section. Next, is a literature survey covering the latest research in this field. The third section explains the research design and the experiments conducted in this study followed by a presentation of the experimental results in section four. Section five gives a detailed analysis of the obtained results. Lastly, the conclusion summarizes the findings of the research and identifies areas of further research.

2. LITERATURE REVIEW

Much of the current literature has investigated the use of Support Vector Machine (SVM), regression trees, linear regression, and deep learning networks to solve the famous regression problem of price prediction, as for instance in [1–3]. Random Forest (RF), a popular tree-based regression technique, was confirmed to have the highest price prediction accuracy in [4]. [3] has shown that linear regression models generally achieve lower accuracy. As opposed to more complex machine learning techniques, like RF for example, which can predict prices very close to the true value but with the cost of a higher time complexity.

The fundamental building block that allows the prediction of product prices using images is Convolution Neural Networks (CNN). In convolution a set of filters is repeatedly applied on an image to obtain a feature map, which in turn shows the location and strength of a feature in the given input. Due to the high capability of CNN to learn a broad number of filters in parallel, distinct features can be identified anywhere on the image.

Hence, this study aims to design a hybrid system that estimates the price of real estate using both numerical and visual data input, as we believe it will achieve high accuracy predictions.

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
Issues (62) 2023
ISSN: 2616-9185
www.mecsj.com

Previous studies have greatly encouraged the use of RF for price prediction, considering its robust and universal performance and its feature extraction property. Another research [4], attempted the prediction of car prices in three different models: model for a certain car make, model for a certain car series and a universal model. The regression techniques used were RF and linear regression. RF showed better performance in the universal model, while the linear regression achieved slightly better results for certain car make and certain car series models compared to RF. Other studies like [5] took a the comparison approach and applied their experiment using the following three models: RF, Gradient Boosting Machine (GBM), and SVM. The findings showed that the tree-based approaches,RF and GBM are more accurate and reliable.

Some studies have examined the compound effect of several machine learning techniques combined together. In [1] for example, RF, SVM, and Artificial Neural Networks (ANN) are tested for price prediction of cars both individually and as a hybrid system. The outcome of the experiment showed that any of the aforementioned techniques achieved less than 50% accuracy if used individually. On the other hand, the ensemble of the the three machine learning techniques scored an average of 87.38%. The strength of hybrid systems is featured in [6], where a combination of GBM and Lasso Regression is used to reach best accuracy. Hybrid systems have the ability to accomplish better results, but with the disadvantage of a more complicated design and a larger time complexity.

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
Issues (62) 2023
ISSN: 2616-9185
www.mecsj.com

Recently, a deep learning model called ”PriceNet” was presented with the purpose of predicting prices of bikes and cars using visual data only [4]. This architecture main body is composed of convolutional and pooling filters and is terminated with a fully connected layer. Compared to Transfer Learning and Linear Regression, which were also tested in this research, ”PriceNet” accuracy exceeds both.

A 2016 research paper [7], proposed CNN model composed of a four node layer and tested it on image and numerical data to prove that a mixed data model can predict prices with higher accuracy than a numerical data only model. Another finding of this paper was that Neural Networks perform better than SVM in price prediction problems. On top of that a new data set was collected, which we used in this study to perform our experiments.

In summary, this literature survey verifies that tree- based approaches like RF and GBM offer the most

accurate and robust price prediction results for numerical and categorical data input. As for price estimation using visual data, CNN should be applied for best results. Thus, this paper aims to investigate the possibility to combine the aforementioned techniques to improve the prediction outcome.

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
Issues (62) 2023
ISSN: 2616-9185
www.mecsj.com

3. METHODOLOGY

3.1 Data Preparation and Analysis

For the first experiment two data sets were tested. The first is the California Houses data set obtained from [7], which consists of more than 500 data points. The second is the Brazilian Houses data set, which is provided by Kaggle [8] and contains around 12 million Brazilian house listings.

Inside the California Houses data set, a house has five numerical attributes and four images, which are:

the number of bedrooms, the number of bathrooms, the area of the house, the zip code of the location, the price of the listing, a frontal view image of the house, an image of the bedroom, an image of the kitchen, and an image of the bathroom. Fortunately, the data set was complete with absolutely no missing values. Nevertheless, the size of the data set was too small for our deep learning application, so we augmented the data set to increase the size to 2000 records. The transformations used were: crop, flip, gaussian blurr, rotate, and scale. The next preparation step was to divide the data into a numerical sub data set and a visual sub data set. The following step was to combine the four images representing each house in one image, as shown in Figure 1. This is important to avoid complexity of the network, which in turn can affect the accuracy of the output. Finally, all images were resized to 224x224 pixels as required by the SqueezeNet model.

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
Issues (62) 2023
ISSN: 2616-9185
www.mecsj.com

In the second data set, on the other hand, each house listing has a total of 24 attributes: id, floors, rooms, created on, collected on, property id, operation, property type, place name, place with parent name, country name, state name, geonames id, currency, description, title, lat lon, lon, lat, surface covered in m2, surface total in m2, expenses, price, house image.

Figure 1:The four images of a house pasted together in a single tiled image.

The number of data records in the second data set is huge, more than 12 million, which is favorable for deep learning experiments. But the data set was incomplete with a lot of missing and redundant values. Given the large volume of data, the first preparation step was choosing a random sample of 5000 data points to represent the entire data set. The next step was to eliminate columns containing text, GPS coordinates, or dates, which were: place name, place with parent name, description, title,lat lon, lon, lat, geonames id,created on, collected on. As these attributes are irrelevant to our study. Next, columns with constant values were dropped(country name, currency and operation). Because these add unnecessary complexity without providing any new information and can bring about unexpected results. Finally, columns and rows with lots of missing values were removed. The final result was a data set consisting of eight features,namely: id, rooms, state name, property type, surface covered in m2, price, and house image. And once more the visual data was separated from the numerical data like in the previous data set.

3.2 Research Approach In our study, we conducted two different experiments. The objective of the first experiment is to test the prediction accuracy of the system for only image input, using the two data sets presented in the previous section. The goal of the second experiment is to test our designed hybrid model estimation accuracy. In order to predict house prices using visual data only in the first experiment, transfer learning technique is used. We trained the SqueezeNet model presented by [9] on both of the data sets, after replacing the output layer with one fully connected layer and a single layer activation output to suit our regression application Figure 2.

Originally, SqueezeNet was designed as a compact alternative for AlexNet presented in [10]. We chose this model for our application, due to the fact that it has about 50 times fewer parameters and is three times faster than AlexNet, while it maintains almost the same accuracy. The main component of the SqueezeNet network is the fire module, which is repeated inside the network.

Adding on the previous experiment, the second experiment creates a hybrid system using the SqueezeNet trained before and RF network. To combine both regression networks statistical methods need to be used. Traditionally, either Mean Predicted Valueor Median Predicted Value , which can be weighted or unweighted, are utilized for this purpose. For gaussian or nearly gaussian distributions the Mean Predicted Value is applied, otherwise if the distribution is unknown the Median Predicted Value is used. Since our model aims to combine only two prediction from the SqueezeNet and the RF networks, we used Mean Predicted Value to calculate the final result of the system, as shown in Figure 3.

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
Issues (62) 2023
ISSN: 2616-9185
www.mecsj.com

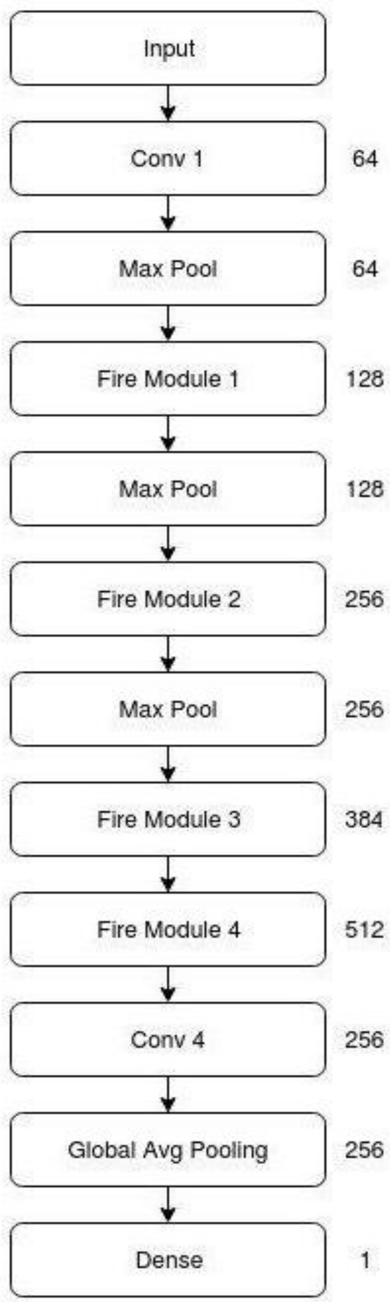

Figure 2:SqueezeNet Flow Diagram.

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
Issues (62) 2023
ISSN: 2616-9185
www.mecsj.com

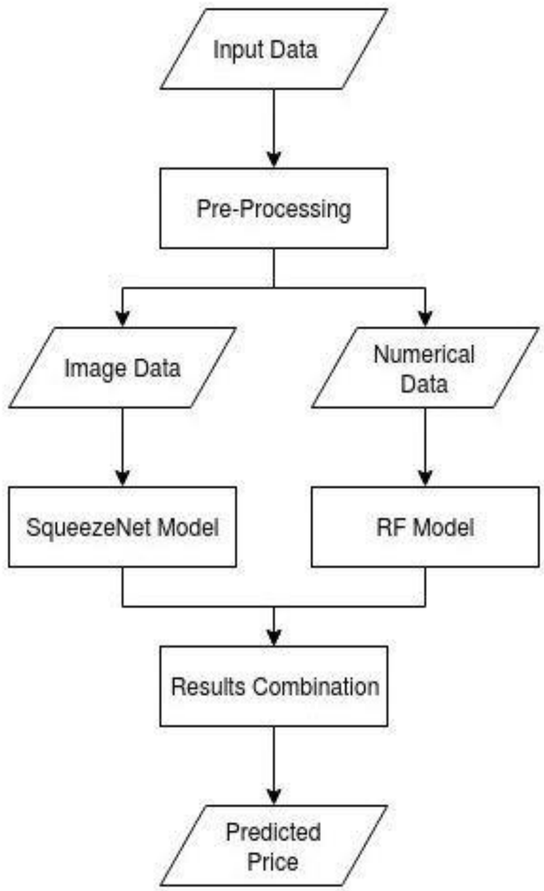

Figure 3: Hybrid System Diagram.

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
Issues (62) 2023
ISSN: 2616-9185
www.mecsj.com

4. RESULTS

4.1 Evaluation Metrics

The choice of a loss function in a Deep Learning experiment is always a very important decision to make. In our experiments, we chose the Mean Squared Logarithmic Error (MSLE) as the loss function as shown by the equation below.

$$L(y, \underline{y}) = \frac{1}{N}\sum_{i=0}^{N} \quad (log(y_{i} + 1) - log(\underline{y_{i}} + 1))^2$$

The main reason for choosing the MSLE, is that the MSLE focuses on the percentual difference between
the log of the actual and predicted values. This allows for a less strict training for a model dealing with large predicted values, as in our case.
On top of that, we calculated the R2 value to give a better perspective of the system performance, as well as to make the comparison with previous models easier and clearer. R2 or the coefficient of determination measures the variance between the true values and the predicted values as in the equation below. R2 values fall between zero and one, where zero means random predictions and one means perfectly correct prediction. In practice, a score of a perfect one is never achieved but the closer the R2 value is to one the better the performance.

www.mecsj.com
Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
Issues (62) 2023
ISSN: 2616-9185

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}$$

4.2 First Experiment

Despite the abundance of data records in the Brazilian Houses data set, it actually resulted in randomly predicted house prices. The MSLE was 0.9, while the R2 was nearly zero. Thus, we did not use this data set in any further experiments. The reasons for this extremely low accuracy to the extend of random results is discussed in detail in the next section of this paper.

On the contrary, the California Houses data set results were rather very good. The MSLE was 0.06 and the R2 was 0.89 during training. During testing the results obtained were MSLE of 0.075 and R2 of 0.78 as in Table 1. If we compare the results of the two data sets used, we see that data quality and preparation plays a crucial role in the accuracy of price prediction using images.

Table 1: Results of first experiment using California Houses data set.

Error	Training	Testing
MSLE	0.04	0.075
R^2	0.89	0.78

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
Issues (62) 2023
ISSN: 2616-9185
www.mecsj.com

4.3 second Experiment

Table 2 below shows the results obtained by the hybrid system presented in this study. It achieved a MSLE of 0.05 and R2 of 0.91, which is impressive for a case as complex as house price prediction. This outcome shows that the SqueezeNet subsystem and the RF subsystem each extracted different features, which explains the increase in accuracy of the joined subsystems in the hybrid model.

Table 2:Results of second experiment..

Error	Squeeze-Net	RF	Hybrid System
MSLE	0.075	0.01	0.05
R^2	0.78	0.86	0.91

5. DISCUSSION

5.1 Analysis of First Experiment

As shown in the previous section, the first experiment using the Brazilian Houses data set yielded a MSLE of 0.9. This high prediction error can be explained by the fact that the images provided by this data set were random, consequently the price predictions were random too. For instance, one record may present the exterior of the house, while another may show an image of the interior of the property Figure 4. It is not specified what part of the property listed, will the image represent. Thus, it was impossible for the SqueezeNet model to learn significant features, that in turn will help the model predict the price during testing. This analysis was confirmed by the fact that the California Houses data set successfully predicted the house values with low prediction error.

Figure 4: Description Is Placed Right Below The Figure

The California Houses data set represents each house with four images. The four mages are of: the

bathroom, the bedroom, the living room, and a frontal view of the house. During data preparation these four images were tiled together in one image in the mentioned order. Therefore, the deep learning network was able to predict the prices with much higher accuracy than it did with the Brazilian data set. The results of this experiment suggest that the model represented by [11] can perform price prediction with high accuracy if the data provided is consistent. Specially, that the original experiment conducted by [11] used two data sets, where all the products were presented in a side view on a white background. Obviously, the Brazilian Houses data set did not fulfill this requirement, which explains the poor performance of the model on this data set and the high accuracy on the California Houses data set.

5.2 Analysis of Second Experiment

Our proposed hybrid system achieved remarkable results as shown in Table 2. The hybrid system has accomplished a higher accuracy than that of the SqueezeNet subsystem or the RF subsystem. This demonstrates that each data type used in each of the subsystems holds different features about the product. so when each data type is trained separately on then both subsystems are combined together a higher accuracy is achieved. The model introduced by [7] has achieved an R-squared of 0.92, while our system has achieved an R-squared value of 0.91 using the same data set,as shown in Table 3. Obviously our model has a slightly lower accuracy, but it has better scalability than the simple CNN presented in [7]. Thus, we believe our model can achieve even better results than the current ones, given a larger data set than the one used in this study.

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
Issues (62) 2023
ISSN: 2616-9185
www.mecsj.com

Table 3: Center Table Captions Above The Tables.

Error	Our Model	Model in [7]
R^2	0.91	0.92

6. CONCLUSION

This paper has argued that it is of great value to include image data in the process of price prediction. And in the case study of real estate listings two experiments were conducted. The first was to test the price prediction accuracy using image data only and the second was to test the accuracy of the proposed hybrid model. The results of this research have proved that the price of houses can actually be predicted from image data with satisfying accuracy. Another finding is that the through preparation of the data is key to obtaining the desired accuracy outcome. Therefore, the limitation of this study is that the data must follow a predefined format as in the California data set. Otherwise, a random set of images will only result in a random price prediction as shown by the Brazilian Houses data set in the first experiment, where the R-squared value was almost zero.

In our work, the results achieved by the proposed hybrid model support the findings of previous literature and prove the importance of the exploitation of image data for better price prediction accuracy. The presented hybrid model successfully achieved a MSLE of 0.05 and a R-squared value of 0.91, which exceeds the accuracy rate of either the RF subsystem or the SqueezeNet subsystem when used singularly.

This research has thrown up many questions in need of further investigation. For instance, the hybrid model could be tested against a data set of a different product. Also, it might be of a great value to research the ability to predict the price of multiple products, which appear simultaneously in the same image. Another way to extend this research would be to try to train the model to recognize the product in any view and still be able to predict the price correctly, which is as mentioned earlier the current limitation of this work. On top of that, to be able to investigate this topic further, it is highly recommended to collect a new large data set suited

for this study. As one of the obstacles faced during this research is finding a proper data set.

Multi-Knowledge Electronic Comprehensive Journal For Education And Science Publications (MECSJ)
Issues (62) 2023
ISSN: 2616-9185
www.mecsj.com

REFERENCES:

[1] Gegic E., Isakovic B., Keco D., Masetic Z., Kevric J.. "Car price prediction using machine learning techniques". TEM Journal, 8(1):113. 2019.

[2] Truong Q., Nguyen M., Dang H., Mei B.. "Housing price prediction via improved machine learning techniques". Procedia Computer Science, 174:433–442, 2020.

[3] Afonso B., Melo L., Oliveira W., Sousa S., Berton L.," Housing prices prediction with a deep learning and random forest ensemble", Anais do XVI Encontro Nacional de Inteligencia Artificial e Computacional, pp. 389–400, 2019.

[4] Chen C., Hao L., Xu C.. "Comparative analysis of used car price evaluation models", AIP Conference Proceedings, vol. 1839, pp. 020165, 2017.

[5] Ho W., Tang BS., Wong S.. "Predicting property prices with machine learning algorithms", Journal of Property Research, vol. 38, pp.48–70, 2021.

[6] Lu S., Li Z., Qin Z., Yang X., Goh R., " A hybrid regression technique for house prices prediction", IEEE international conference on industrial engineering and engineering management (IEEM), pp. 319–323, 2017.

[7] Ahmed E., Moustafa M.. "House price estimation from visual and textual features". arXiv preprint arXiv:1609.08399, 2016.

[8] Data science challenge at EEF 2019 [Internet]. Kaggle.com. [cited 2022 Sep 27]. Available from: https://www.kaggle.com/c/data-science-challenge-at-eef-2019.

[9] Iandola F., Han S., Moskewicz M., Ashraf K., Dally W., Keutzer K.. “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size”. arXiv preprint arXiv:1602.07360, 2016.

[10] Krizhevsky A., Sutskever I., Hinton G.. “Imagenet classification with deep convolutional neural networks”. Advances in neural information processing systems, vol. 60, pp. 84-90, 2017.

[11] Yang R., Chen S., Chou E.. “AI blue book: vehicle price prediction using visual features”. arXiv preprint arXiv:1803.11227 2018.