STM JOURNALS

Recent Trends in
Parallel Computing

https://journals.stmjournals.com/rtpc

ISSN: 2393-8749
Volume 11, Issue 1, 2024
January—April
DOI (Journal): 10.37591/RTPC

Review RTPC

Interest Level Prediction in Rental Properties Using Data Science

V.R. Siva[1,*], R. Durga[2]

Abstract

A key component of forecasting home prices and rental patterns is real estate market analysis. Data science, data mining methodologies, and statistical models are some of the strategies that have been created in recent years to solve this problem. A few problems are still required to be resolved, such as the obstacles caused by the availability and quality of the data; the presence of outliers, missing values, and inconsistent formats in housing datasets might have a negative impact on prediction models' performance; and ensuring data accuracy and coverage necessitates working with government agencies and data providers in conjunction with meticulous data preparation processes. A number of major issues impede the accuracy and dependability of current systems, like only a little amount of thorough and current data is readily available, market dynamics are complicated, and effective.

Keywords: Data science, forecasting home prices, real estate market analysis, data mining methodologies, key component, rental patterns, market analysis

INTRODUCTION

It sorts and displays the listings to users using a creative algorithm. We took part in the Kaggle machine learning competition. Anticipating the level of interest in an apartment rental listing was our difficulty. We predicted 49,000 out of the 74,000 labeled entries that we had. The target designations were medium, low, and high [1].

This problem's scope is representative of a common machine learning difficulty encountered by numerous firms. Missing values, outliers, and inconsistent formats can have a negative impact on how well prediction models function for housing datasets. It included geographical data, unit descriptions, and images. There was also an imbalance in the classes' distribution. Considering all the factors, we can confidently state that this problem is a far better proxy [2].

IMPLEMENTATION

To have an exceptionally computerized and controlled climate for our elements, where we guarantee that preparation and test information go through similar changes from crude information to becoming contributions for our brain organizations, we fostered a preprocessing system, with numerous potential changes, that follows through on that commitment. In the wake of setting up the preprocessor, with every one of the various pipelines for the various sorts of information handled, getting the information is just about as basic as calling load and transform(test), with test being False for train and True for test information [3].

With the generator module, this system stretched out to information stacking is lined up as the

***Author for Correspondence**
V.R. Siva
E-mail: rajendransiva00@gmail.com

[1]Student, Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies, Pallavaram, Chennai, Tamil Nadu, India
[2]Associate Professor, Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies, Pallavaram, Chennai, Tamil Nadu, India

Received Date: February 16, 2024
Accepted Date: March 02, 2024
Published Date: April 04, 2024

Citation: V.R. Siva, R. Durga. Interest Level Prediction in Rental Properties Using Data Science. Recent Trends in Parallel Computing. 2024; 11(1): 28–34p.

organization is being prepared. This is a basic usefulness when all the photograph information, once stacked, would surpass the memory limit of our framework [4].

Initial Approach

To begin solving this problem, we first divided and sliced the provided variables in different ways in order to extract information from them. The almost readily available ones, that did not require much treatment, are the following:

- *Price (continuous):* Log transformation.
- *Bedrooms (integer):* Unchanged.
- *Bathrooms (integer):* Unchanged.
- *Value/Rooms and Value/Washrooms:* Utilized by adding 1 to the denominator and taking the logarithm [5].
- *Scope and Longitude (nonstop):* After an underlying investigation, it was resolved that, regardless of not having all directions highlighting the New York City region, that the ones that did not were generally wrong and alluded to postings from New York City. Subsequently, we constrained all directions into a rectangular region around the city.
- *Photographs (rundown of URLs):* As a first methodology, utilized basically the number of photographs for each posting.
- *Portrayal and Highlights (text, list):* As a first methodology, separate straightforward measurements like the length and number of words.
- *Uploading the creation date and time:* Clearly marked highlights such as the day, hour, and month.

Price

These underlying highlights were at that point exceptionally educational in that it was feasible to acquire cross-entropy misfortunes underneath 0.6 (top score is still above 0.500 as we compose this article) with specific models. The price was found to be the most important of these features [6]. Plotting the cost thickness for various interest levels. The scale of price is logarithmic as shown in Figure 1.

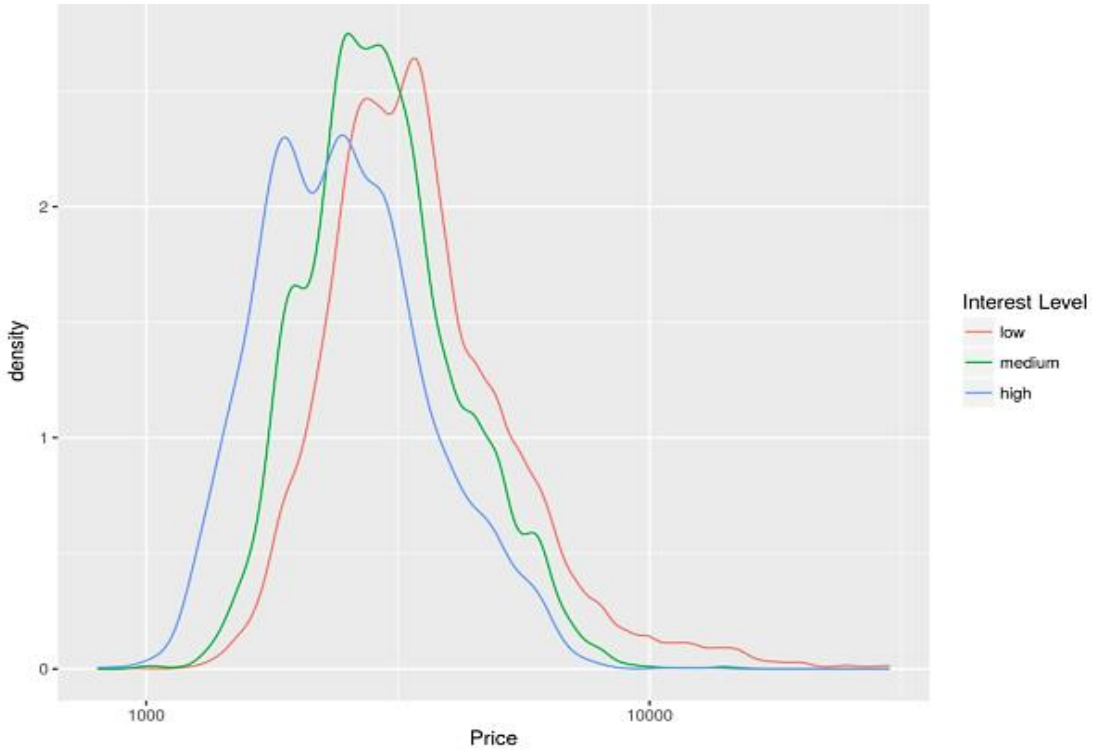

Figure 1. Scale of price.

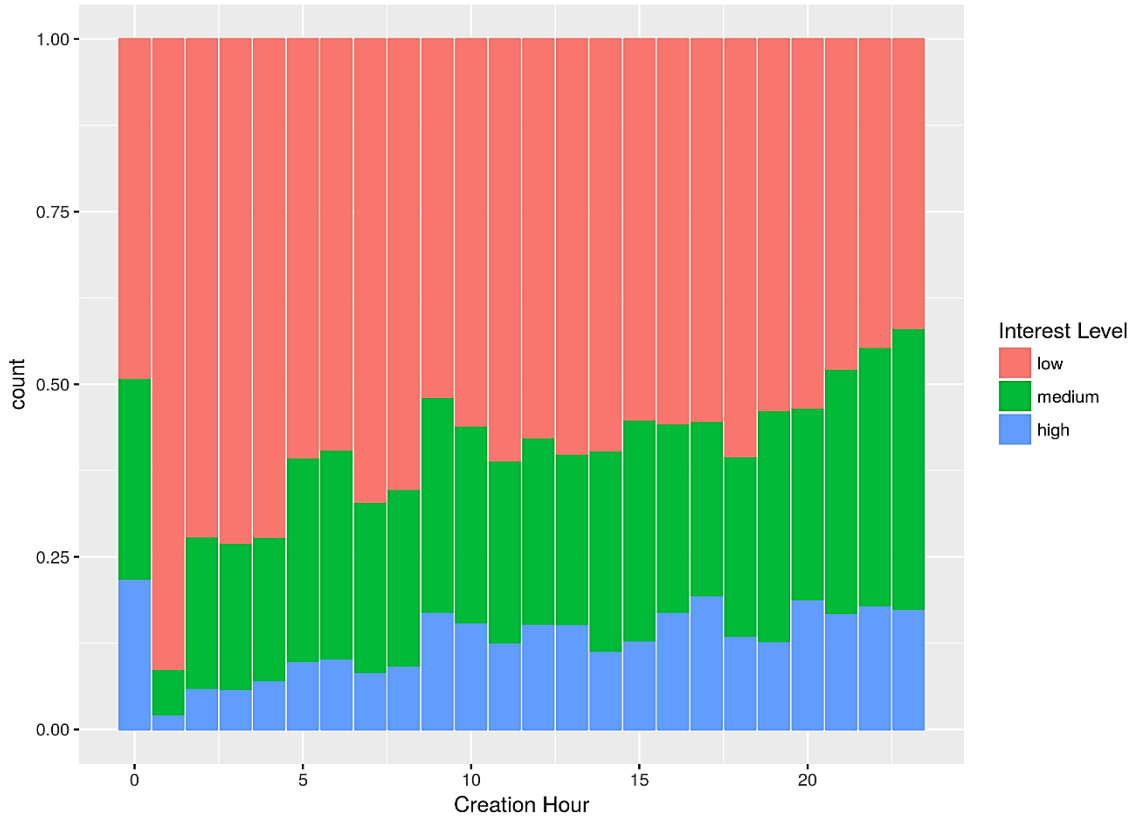

Figure 2. Interest levels by listing creation hour.

Interest Level

Additionally, the listing creation hour may provide insight into the degree of interest as shown in Figure 2.

Other All-out Factors

A methodology in some cases utilized on all out factors comprises encoding them (changing to number qualities) and changing them into sham factors. The indicators being referred to are Director ID, Building ID, Road Address and Show Address [7].

Posting Elements

With regards to utilizing loft highlights as an indicator, we needed to begin by looking hard and long at the information. On our crude information, the posting highlights were either introduced in an extremely organized manner, like in Elevator, Clothing in Building, Hardwood Floors; or in an exceptionally messy way, for example, in **LIFE OF Extravagance FOR NO Charge!* Rambling 2BR/2BA Chateau *WALLS OF WINDOWS*;

Ample Wardrobes *FREE Exercise center and POOL* Grand Rooftop DECK *Custodian/ELEV BLDG*;

Moves toward THE Recreation area!! **.

With this information, we decided that regular expressions were the best way to capture features, making sure that all matches were relevant to the intended feature. In total, we extracted 56 distinct listing features [8]. Out of the highlights removed, the ones that appear to have the best effect as far as the interest levels, thinking about how much postings they influence, are Hardwood Floors and No Expense as shown in Figures 3 and 4.

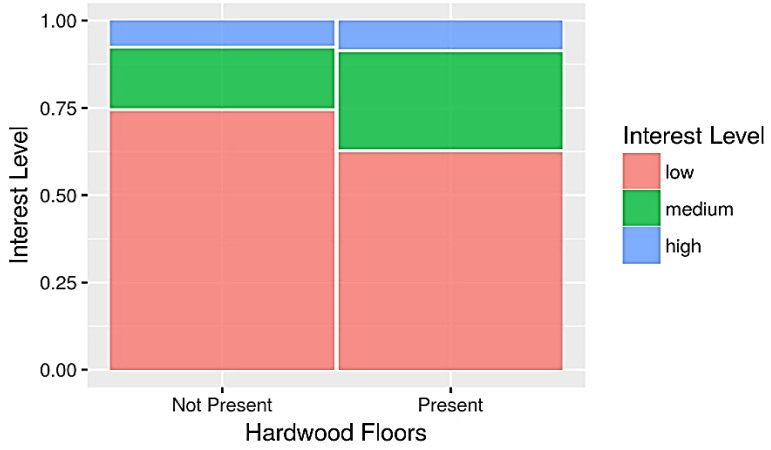

Figure 3. Distribution of interest levels with respect to hardwood floors listings.

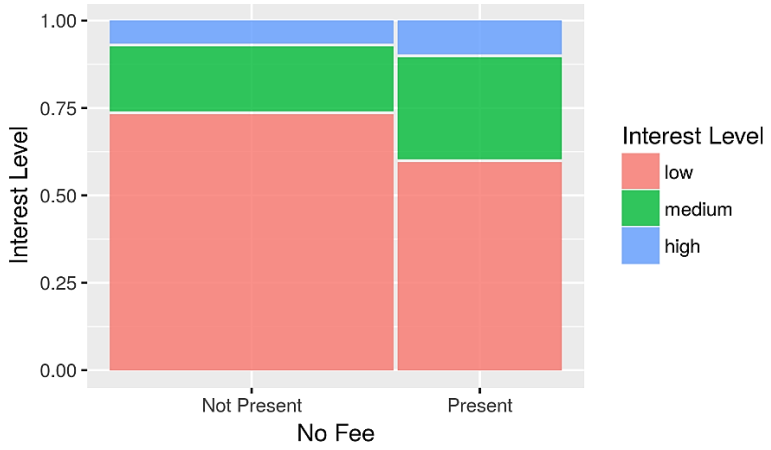

Figure 4. Distribution of interest levels with respect to no fee listings.

PROPOSED WORK

In view of the way that the vast majority of the information, regarding size, made accessible to us was as photographs, we needed to utilize it on our expectations. The test here is to separate from them a larger number of highlights than the promptly accessible ones: number of photographs per posting and their aspects [9].

We recognized four potential methodologies (not totally unrelated):

- Highlight Designing: remove a few physically picked insights, like splendor, sharpness, contrast, and so forth.
- Create a distinct convolutional neural network model and train it to categorize the images according to what they show. Potential classifications could incorporate kitchen, washroom, floor plan, wellness focus, road view, and so on.
- Train a different model with just the photographs, then, at that point, feed the outcomes to our principal model.
- Remove unknown highlights via preparing on a bound together model.

At first, as a planning move toward carrying out the last methodology, we prepared a straightforward brain network model with four thick layers, taking as info the fundamental posting highlights previously furnished to us with only a couple of changes, and continuously fostered our element designing and saw our expectations get to the next level as shown in Figure 5.

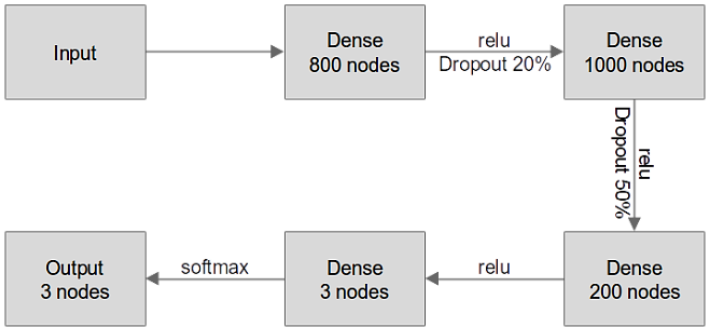

Figure 5. Simple neural network model.

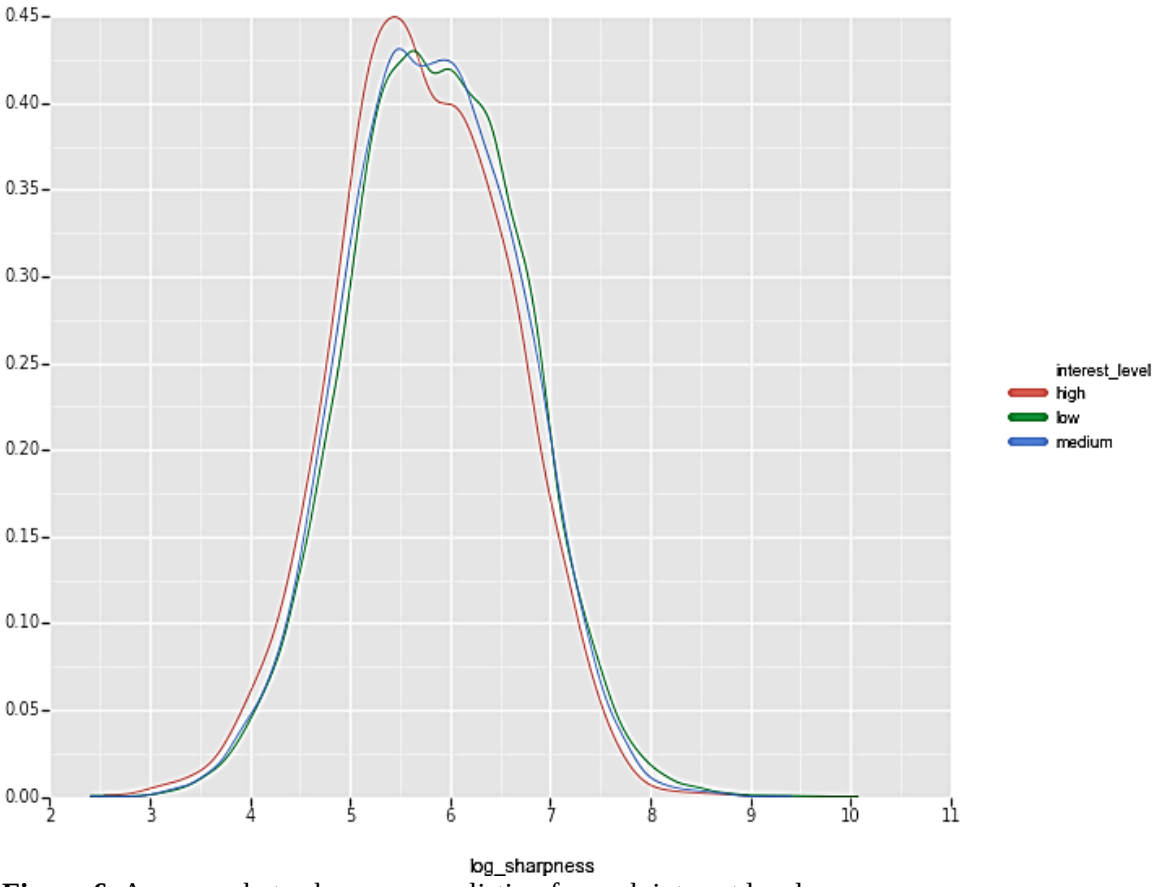

Figure 6. Average photo sharpness per listing for each interest level.

Extracting Data

In the wake of getting all the photographs, taking altogether over 80 GB of extra room, we chose to separate a couple of helpful qualities from them [10].

In the first place, we viewed at their aspects and remembered them as indicators for our models (specifically, normal aspects per posting, as well as the components of the principal photograph).

Additionally, we extracted the sharpness of each image, resulting in a somewhat unexpected conclusion: in opposition to our underlying presumptions, exorbitant interest postings have somewhat less sharp photographs when contrasted with low interest postings as shown in Figure 6.

Sharpness of photos is typical for each interest level per posting. We likewise utilized Clarifai Programming interface to remove top 15 names from each picture. There, convolutional brain networks

are utilized to learn highlights present in the picture and a likelihood gauge is given for each name separated. However, under inclination supporting relative impact, picture highlights showed some certain importance, during preparing and testing they gave exceptionally low indications of working on the model.

At long last, to really utilize the photograph contents, we resized all photographs to 100×100 squares, to be taken care of into a convolutional brain network model (more subtleties on that underneath).

Feature Transformation

Feature Transformations: In the end, the model with the highest Kaggle score, 0.58854, produced the best results. It used the following feature transformations:

- Facilitates (longitude/scope) outside the New York City region snapped to a square shape around it,
- Logarithm of cost, cost per room and cost per restroom,
- Count of words/characters in depiction, words and amount of condo highlights,
- Season of day, day of month, day of week,
- Opinion investigation,
- Parsed condo highlights as 56 sham factors,
- Aspects and sharpness (log) of first photograph, as well as normal aspect and sharpness (log) of all the photographs per posting,
- Administrator id: encoded the main 999 directors concerning how much postings, planning every one of the leftover ones to a typical classification, and afterward applied an implanting to 10 initiations. Like creating 1000 faker factors and afterward applying a thick layer with 10 enactments as shown in Figure 7.

We encountered a few challenges when attempting to use the original photographs. They varied widely in terms of aspect ratios and sizes, to start. The second thing was that the quantity of images varied throughout listings. Lastly, storage limitations with the GPU servers we employed. We were able to fit them all into approximately 2.3 GB of storage by resizing them all to 100×100 square thumbnails.

We chose to use only the first photo as a first approach and to only take 20,000 training samples in order to address the fact that different listings had varying numbers of photos and the memory space inflated photos take as inputs to a convolutional network (100×100×3×4~120 kb) as shown in Figure 8.

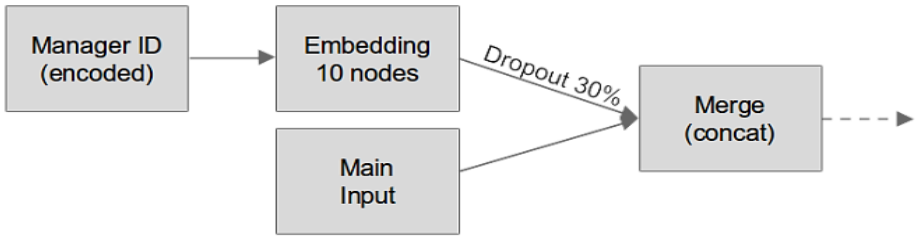

Figure 7. Neural network embedding for Manager ID.

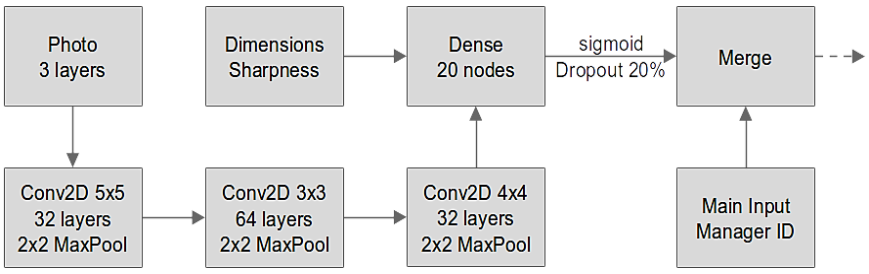

Figure 8. Convolutional neural network channels.

CONCLUSION

After a few hours of training, the validation set's results did not seem all that promising. We made the decision to extract the activations from the convolutional portion of the network, freezing its (trained) weights and eliminating the need for the photos for training, and to expand, up to validation data, to the entire training set, since high-quality training data is critical to a high-quality predictor.

We did see some improvement from this, but strangely not enough to surpass our prior Kaggle score that did not include the photographs. This seems to contradict the basic intuitive notion that, if a model is trained well, it should be able to make better predictions given more data and freedom (weights). To be honest, virtually none.

REFERENCES

1. Bergstrom CT, West JD. Calling bullshit: The art of skepticism in a data-driven world. Canada: Random House Trade Paperbacks; 2021 Apr 20.
2. Chang W. R graphics cookbook: practical recipes for visualizing data. USA: O'Reilly Media; 2018 Oct 25.
3. Cleveland WS, McGill R. Graphical perception and graphical methods for analyzing scientific data. Science. 1985 Aug 30; 229(4716): 828–33.
4. Yau N. Visualize this: the FlowingData guide to design, visualization, and statistics. John Wiley & Sons; New Jersey, United States. 2011 Jun 13.
5. Wickham H, Grolemund G. R for data science: Import, tidy, transform, visualize, and model data. USA: O'Reilly Media, Inc.; 2017.
6. Wickham H. ggplot2: Elegant graphics for data analysis. 2nd Edn. New York: Springer-Verlag; 2016.
7. Pearl J, Mackenzie D. The book of why: The new science of cause and effect. New York City: Basic Books; 2018.
8. Neth H. ds4psy: Data science for psychologists. Social Psychology; Decision Sciences. Germany: University of Konstanz; 2020.
9. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013 Sep 1.
10. Foreman JL, Gubbins EJ. Teachers see what ability scores cannot: predicting student performance with challenging mathematics. J Adv Acad. 2015; 26(1): 5–23.

