JREFE Revisions

Author Response

May 31, 2019

Response to Referee Comments

Referee 2

This paper estimates the value of curb appeal in SF house prices. A classification analysis is used to score curb appeal based on Google Street View images. The score is then used, along with a number of other hedonic features, to predict home values. From the results one can infer the value of curb appeal. I find the analysis and results interesting. I'd recommend publication after a revision that addresses the following issues:

Comments

Details on scoring

More details need to be provided on the curb appeal scoring. For example, how many homes were scored manually to create the training set for the classifier? The authors need to provide all details to enable replication.

In the paper given to the referee, we used 50 photos of each curb appeal class to the build the model. In response to previous referee comments and conference feedback, we have updated the model to use 100 training photos from each class. This improved our out of sample fit greatly and clarified some of the results where we examine the hedonic effects by neighborhood quality type. We return to this discussion below.

We also provide links to all data, code and trained models that would be needed to replicate and verify our model.

The model, labels, python code and classified photos are saved on our publically accesible github page at https://github.com/erikbjohn/curb_appeal/tree/master/Replication. We provide a link to the necessary replication files in the paper. Importantly, this includes all code and data needed to replicate the training and model creation as well as the model itsself.

Transfer Learning Classifier

How well does the transfer learning based classifier perform in- and out-of-sample (this could be measured by 2-class AUCs)? For out-of-sample testing the authors need to construct a manually labeled test set. The performance of the classifier needs to be established before it can be used.

We include the following analysis in the paper and below.

In sample analysis

First, we start with the out of sample classifier accuracy by manually labeling a set of photos. We selected photos until we had 100 out of sample photos in each of the 4 classifications. We start by comparing the 'ground-truth' scores that we manually labeled to their highest probability scores. This analysis is done using a confusion matrix that allows to illustrate the relationship between the Predicted and Reference estimates in Table 1. Overall statistics are shown in Table 2.

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

	Reference: 1	Reference: 2	Reference: 3	Reference: 4
Prediction: 1	93	8	0	0
Prediction: 2	6	85	7	5
Prediction: 3	1	4	85	8
Prediction: 4	0	3	8	87

Table 1: In sample confusion Matrix

Statistic	Value.V1
Accuracy	0.88
Kappa	0.83
AccuracyPValue	0.00

Table 2: Overall Statisitics

Class diagnostics are provided in Table 3 and in the manuscript. Discussion of the diagnostics may be found at https://topepo.github.io/caret/measuring-performance.html. Importantly, Precision is the same as size and Recall corresponds to Power.

	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.93	0.85	0.85	0.87
Specificity	0.97	0.94	0.96	0.96
Pos Pred Value	0.92	0.83	0.87	0.89
Neg Pred Value	0.98	0.95	0.95	0.96
Precision	0.92	0.83	0.87	0.89
Recall	0.93	0.85	0.85	0.87
F1	0.93	0.84	0.86	0.88
Prevalence	0.25	0.25	0.25	0.25
Detect Rate	0.23	0.21	0.21	0.22
Detect Prevalence	0.25	0.26	0.24	0.24
Balanced Accuracy	0.95	0.90	0.90	0.92

Table 3: In sample Diagnostics by class

Out of sample sample analysis

We next examine out of sample analysis by manually scoring apporoximately 100 photos into each of the categories and compare the model predictions. The diagnostic statistics are similar to the in-sample training analysis, but with a bit less accuracy as is to be expected.

Overall statistics are shown in Table 4.

Class diagnostics are provided in Table 6 and in the manuscript. Discussion of the diagnostics may be found at https://topepo.github.io/caret/measuring-performance.html. Importantly, Precision is the same as size and Recall corresponds to Power.

Class diagnostics are provided in Table 3 and in the manuscript. Discussion of the diagnostics may be found at https://topepo.github.io/caret/measuring-performance.html. Importantly, Precision is the same as size and Recall corresponds to Power.

Statistic	Value.V1
Accuracy	0.66
Kappa	0.52
AccuracyPValue	0.00

Table 4: Overall Statisitics

	Reference: 1	Reference: 2	Reference: 3	Reference: 4
Prediction: 1	7	5	4	0
Prediction: 2	2	16	7	0
Prediction: 3	0	4	25	7
Prediction: 4	0	1	3	15

Table 5: Out of sample confusion Matrix

	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.78	0.62	0.64	0.68
Specificity	0.90	0.87	0.81	0.95
Pos Pred Value	0.44	0.64	0.69	0.79
Neg Pred Value	0.97	0.86	0.77	0.91
Precision	0.44	0.64	0.69	0.79
Recall	0.78	0.62	0.64	0.68
F1	0.56	0.63	0.67	0.73
Prevalence	0.09	0.27	0.41	0.23
Detect Rate	0.07	0.17	0.26	0.16
Detect Prevalence	0.17	0.26	0.38	0.20
Balanced Accuracy	0.84	0.74	0.72	0.81

Table 6: Out of sample Diagnostics by class

Out of sample performance - Regression

It would also be good to see out-of-sample tests of predictive performance for the actual price regression model.

Referee Number 1

Summary of the Paper

This paper uses a machine learning (ML) technique to quantify how much curb appeal accounts for house price in a hedonic model. This is an interesting application of ML to housing research. Curb appeal is a potentially important factor that affects home prices but is usually difficult to measure. I agree with the authors that the use of ML to quantify how important curb appeal is an improvement in measuring curb appeal over other methods such as survey-based ones.

In the paper, ML enables the authors to assign curb appeal scores by examining photos of an exterior of a house obtained from Google Street View. Then the score is included as an independent variable in a hedonic regression of house prices. The authors find a modest effect of curb appeal on house prices: one standard deviation increase in the own curb appeal score results in a 1.2% increase in the house price. They find that across-street neighbor's score also matters with one standard deviation increase in the score leading to a 0.6% increase in the house price. The authors also explore whether the effect of curb appeal is heterogeneous depending on housing market cycles and neighborhood-level curb appeal.

Overall, the use of ML could be a potentially important contribution to the literature. However, I am afraid that the paper's contribution is quite limited. My detailed comments about the paper follow in next sections.

Comments

What is captured by the score?

My main complaint about this paper is that the authors' method seems to measure only limited aspects of curb appeal. One can easily imagine that there are hundreds of attributes that describe curb appeal of a property such as the architectural style of a property, the color of an exterior, types of trees in the front yard, etc. However, it is not very clear how these attributes were taken into account by the authors' scoring. Given how the authors describe how they assigned curb appeal scores, only a very limited set of such attributes are likely to be captured. I am concerned that the authors' curb appeal scoring underestimates how much curb appeal matters because it does not capture lots of attributes that determine curb appeal. Indeed, the authors find a modest effect of curb appeal: one standard deviation increase in the own curb appeal score results in a 1.2% increase in the house price. This estimate is much smaller than the estimate by Glaeser et al (2018), who use ML (Convoluted Neural Network) to generate 100 vectors that capture attributes of an exterior of a house and directly relate them to house prices. They find that one standard deviation improvement in house appearance increases the house value by 16%, which is larger than the authors' estimate by more than ten times.

Subjective Evaluation

The authors use an ML technique to assign curb appeal scores automatically to properties by having an algorithm to examine a photo of an exterior of each property. A critical step required for this procedure is the initial subjective evaluation of curb appeal of properties in the randomly selected training sample. The authors manually assigned discrete scores ranging from one to four by looking at photos of properties in the training sample.

I am afraid that this manual assignment/ subject evaluation critically limits this paper's contribution. There will always be cases for which one cannot clearly decide which score to assign, and any associated measurement errors may result in attenuation bias. Moreover, why are there only four categories? This seems quite arbitrary, and I am wondering that the four categories capture curb appeal only in a limited way, which again may result in underestimating how much curb appeal matters. I can also imagine that if the authors increase *the number of possible categories, it will be increasingly difficult to manually determine curb appeal scores. If one were to allow for ten categories, for example, it will be probably very difficult to manually distinguish which property should get 9 vs 10. So, I am afraid that the authors' algorithm is not flexible enough to allow even small tweaks in their procedure.*

In addition, the fact that the authors' method involves subjective evaluation likely makes it difficult to apply the author's method to different settings. For example, the scoring for Denver might be difficult to transfer to homes in Arizona, whose exteriors probably have different styles. For other settings, one would probably need to manually assign curb appeal scores again for the authors' method to work, which makes it costly to use the authors' method.

Contribution

Lots of aforementioned issues could have been addressed if the authors had fully utilized the strength of ML. In general, an advantage of ML is that researchers can be agnostic about which attributes are important a priori and let an algorithm determine that. In fact, because how these attributes matter for property value can be very subjective, it is inherently difficult to evaluate which properties have more curb appeal for actual home buyers, not the researchers. Moreover, if curb appeal is inherently continuous, any discrete scoring may lead to measurement errors. An ideal of use of ML would be to create an extensive list of property attributes from property images (as in Glaeser et al, 2018) and let an ML algorithm determine how much they matter for home prices.

Then it is difficult to see what this paper's contribution is relative to Glaeser et al (2018). The ways in which the two papers estimate how much curb appeal accounts for home prices are different. But the authors' method seems inferior.

One way to differentiate the authors' paper from Glaeser et al (2018) is through the heterogeneous effects. However, the authors do not provide a reason why looking at these heterogeneous effects are important and interesting. Are these heterogeneous effects supposed to test predictions of a theoretical model? Providing a more coherent story around these heterogeneous effects would be helpful for a reader to see a contribution of this paper.

Unobservable Attributes

An identifying assumption for the hedonic regression is that there are no unobserved home attributes that are correlated with the curb appeal score. The authors might want to discuss how plausible this assumption is in their setting. An example of possible violation of the assumption is a distressed sale. In that case, houses would be sold at a discount, and such properties would have low curb appeal scores because the foreclosed properties would not be maintained very well. In fact, the sample period is from 2008–2018. Thus, foreclosed properties are likely included in the sample. Moreover, the finding that curb appeal matters more in a cold market (2008–2012) is also consistent with the possibility that the estimated coefficient for curb appeal may reflect a distressed sale at least to some extents. So the authors may want to see how their results change after excluding foreclosed properties from the sample.

How much does curb appeal account for home prices?

One of main motivations to include curb appeal is to address omitted variable bias in a typical hedonic regression, which does not include curb appeal in the right hand side. However, the authors do not show how much including curb appeal in a regression changes the result or how serious the omitted variable bias would be without curb appeal. Can you show results without curb appeal scores? You can also show how much does R-squared increase with curb appeal scores.

Other Minor Comments

Is it possible to allow for an interaction term between own and neighbor's curb appeal? You find that own curb appeal becomes more important if the average neighborhood-level curb appeal score is higher. I am wondering whether this interaction also exists at a more finer level.

You find that across-neighbor's curb appeal does not become more important in a cold market, whereas own curb appeal becomes more important. Can you explain possible stories for why?