Review of "Deciphering House Price Dynamics Using Predictive Modeling"

Reviewers: **Ian Jackson**, Rachel Lim, Simon Fei, Qisen Yang

Zou et al. explore the Ames Housing dataset, proposing a decision tree model for predicting housing prices from a subset of available features. They first use a combination of forward selection, k-fold cross-validation, and grid-search to select maximally predictive features in the large dataset. After finding a 12-parameter model with minimal mean absolute error and mean-squared error, authors train their decision tree on the preprocessed dataset. Using this model, authors ultimately predict housing prices from the 12 features, with the square footage of the open porch, the existence of an above-grade living area, and the square footage of the basement being among the most statistically significant features.

Overall, Zou et al. present a useful investigation of the predictability of housing prices based on just a few key factors. Before accepting for publication, however, some concerns with this manuscript must be addressed:

Major Concerns:

1. Clarity of hypothesis

Authors state their hypothesis as the following:

"...we hypothesize that certain factors such as property type, location, lot size, overall quality and condition will have a significant impact on house prices. Additionally, we anticipate that the use of forward selection, cross-validation, and grid-search to select a decision tree model, will enable accurate estimations of house prices."

This hypothesis needs more clarity and definition, and begs several questions: 1) What *exactly* were the factors expected to have an effect on the house prices? 2) What were the effects expected for each factor (i.e., were you expecting a positive relationship, or otherwise? In other words, it is not sufficient for a hypothesis to list only a few of the expected affecting factors without also outlining the nature of the effect.

Furthermore, the combination of feature selection, cross-validation, and grid-search are useful for observing the hypothesized effects, but not in themselves experimental. The desired effect of increased accuracy is implicit in the use of those techniques. Authors should instead list the above as techniques for model selection, rather than a separate

hypothesis. Alternatively, the result of *not* using the above techniques should be investigated in order to claim any influential outcome.

2. Related studies and background

The study (especially the introduction section) is lacking in exploration of related research, which is key in characterizing the question they seek to investigate in a scientific context. What kind of data has been explored previously? What methods were used? See Mohd et al. (2020), Geerts & de Weert (2023), and Bafna et al. (2018), and others for recent reviews of predicting housing prices using machine learning techniques. This could also aid in hypothesizing which features might be more predictive of housing prices.

3. “Importance value” and interpretation of results

In the Results section, authors mention an “importance value,” which was used to determine the statistical impact of each feature used in the prediction. Furthermore, the paper includes a table of the importance values for each feature. Although this appears to be a useful metric, no quantitative definition for this metric is provided. Be it a probability metric for decision trees, or another statistical measure, it is important to define the meaning (and calculation) of the metric upon its introduction in the section.

Furthemore, the interpretation of the results is lacking in qualitative discussion. Why might the observed features be more useful for prediction than others? A discussion of this can help reconnect the results to the original question at hand.

4. Model comparison needed

The assignment given asks that some kind of model comparison must be done. This is especially useful for contextualizing how well the proposed decision tree model works (i.e., providing a comparison MSE/MAE metric). We suggest comparing the decision tree technique with a simple multiple linear regression, either in this iteration of the project or in future work if time does not allow.

5. Figure 1

The purpose and function of Figure 1 is confusing. It is presented as showing the ranges of hyper parameters used for the grid search. However, it is stated in the paper that only some of the hyperparameter ranges were used while others (min_samples_split and min_impurity_decrease) were constant. Because the grid

search parameter sweep is determined by the researchers, it is confusing why this occurred. Was it a programming error, or something else? It may be sufficient to simply state the minimum and maximum values, and the step size, for each parameter in the search. Because each value searched per parameter is an even distribution, the box plots don't serve a clear purpose for visualizing them.

Minor Concerns:

1. Data preprocessing

Authors implemented several data preprocessing steps, which led to a severe decrease in the amount of data available (around half of the data was removed due to NaN values). Though a reasonable approach, authors should make more considerations before removing data completely, including: what were the proportion of NaN values for each feature? Did one feature have more NaN values than another? With answers to these questions, it could be a viable alternative to interpolate sparser NaN values (i.e., those with more available surrounding data to use for interpolation).

In addition, it is also a useful consideration to investigate the number of repeat houses in the dataset (if any, at all). The presence of multiple of the same house IDs could skew the data, though this was not mentioned in the paper. It may be useful to explore different techniques for handling repeat instances, for example using the average of the values or using only unique instances.

2. Units for MSE and MAE

The units for the model evaluation metrics are unclear. They are assumed to be in dollars ($USD), but should be stated clearly in the paper.

3. Dataset exploration and visualization

Some data exploration is left to be desired in order to convey to readers the nature of the dataset. In other words, what is the underlying structure of the data? Are there any strong correlations between predictors? These questions are related to the aforementioned concerns of data preprocessing, and some visualization of the original data (e.g., mean values and standard deviations) would be helpful.