Cover Letter

In response to the feedback received, we have made several substantive changes to our data preprocessing and model evaluation approaches to address reviewers’ concerns. Firstly, we re-evaluated our initial strategy of removing rows with missing values. Instead, we implemented advanced imputation techniques using scikit-learn’s IterativeImputer for numerical columns and SimpleImputer for categorical columns. This allowed us to retain significantly more data, preserving the dataset's integrity and improving model robustness. We carefully reviewed the proportion of NaN values for each feature, pruning only columns with more than 20% missing values, resulting in the removal of six columns out of seventy-nine. Furthermore, we confirmed that each entry had a unique house ID, ensuring there were no repeated entries. These adjustments did cause significant changes to the results, so we adjusted the writeup accordingly.

Regarding the critique of our choice of k-fold cross-validation, we adjusted our approach from 10-fold to 5-fold cross-validation. This change increased the test set size for each fold, enhancing the reliability of our model evaluation. While several reviewers suggested incorporating alternate models for comparison, we focused on optimizing a single decision tree model in line with guidance from our professor. Specifically, we received concerns from Professor Eran Mukamel about incorporating three different models at one point (linear regression, decision tree, and support vector machine), so we stuck with a decision tree for this project. To address concerns about overfitting, we expanded our discussion on how k-fold cross-validation and grid search techniques help mitigate this issue. Some have suggested adding more descriptions to the variables themselves, but we instead provided a way to access to the data\_description.txt file in the dataset, which contains all the descriptions of the variables. This way, it wouldn’t clutter the writeup itself. Additionally, we acknowledged the limitations of using the Ames Housing dataset, discussing potential impacts on generalizability and suggesting future research directions to validate our findings with diverse datasets. We refined our hypothesis to clearly outline expected relationships between key factors and house prices, and we added explicit definitions for critical variables to improve reader understanding. We added an additional related works section, visual clarity of figures (mostly increasing font size), and more detailed captions. Lastly, we added a detailed explanation of the importance values metric.

Explain how you have responded to each of the reviewers’ major critiques

There were a number of people who’ve expressed concern about pruning a significant amount of data during the pre-processing step.

* **Uditi Namdev**: “For the data preprocessing part, did you try imputing the missing values using any advanced techniques? Removing rows should ideally be the last resort.”
* **Jemin Vagadia**: “The approach to handling missing data by pruning columns with a substantial proportion of NaN values and removing rows with remaining NaN values could introduce bias. A more sophisticated imputation method could be considered to preserve the integrity of the dataset and improve the robustness of the predictive model.”
* **Ian Jackson**: “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.”
* **Yanghong Wei**: “The study mentions removing features and rows with significant missing values, leading to a loss of over half the dataset. This reduction might have impacted the robustness and generalizability of the model. As a potential solution, the report could benefit from exploring advanced imputation techniques to handle missing data more effectively.”

In response to the feedback, we made several significant changes to our data preprocessing approach to address the concerns raised.

Firstly, we reviewed the proportion of NaN values for each feature. Columns with a substantial proportion of NaN values (approximately 20% or more) were still pruned as they could introduce significant bias and instability. This step was carefully considered to ensure that only the most problematic columns were removed. There were only 6 columns out of 79 removed: LotFrontage, Alley, FireplaceQu, PoolQC, Fence, and MiscFeature.

Next, we replaced the initial approach of pruning rows with missing values with imputation methods. Specifically, we employed scikit-learn’s IterativeImputer for numerical columns and SimpleImputer for categorical columns. The IterativeImputer models each feature with missing values as a function of other features, which allows for more accurate imputation. The SimpleImputer, set to most\_frequent, replaces missing categorical values with the mode of each column. This method retained significantly more data (over half the data) than our previous approach of just pruning all the rows that has missing entries.

Additionally, we investigated the presence of multiple instances of the same house IDs in the dataset. We found that each entry had a unique ID, thus there were no repeated houses in the dataset.

These changes did result in significant changes to the results section, so they’ve been modified accordingly too. In fact, we got noticeably better results with imputation, so we thank the reviewers for bringing this to our attention.

One person brought up a concern about the choice of k-fold cross-validation.

* **Uditi Namdev**: “The decision tree modeling approach and forward feature selection method seem well-justified. However, with k=10 for cross-validation, the test set size becomes quite small for each fold, which could impact the reliability of model evaluation. I’m assuming that the grid search process helped mitigate this issue by exploring different hyperparameter configurations extensively.”

Specifically, she noted that with k=10, the test set size for each fold becomes quite small, potentially impacting the reliability of model evaluation. To address this issue, we have adjusted our cross-validation approach to use 5-fold cross-validation instead. This adjustment increases the size of the test set for each fold. With 2,335 data entries in our dataset, this change results in each fold containing approximately 467 data entries for validation and 1,868 data entries for training.

There were also concerns about incorporating additional models into the project.

* **Uditi Namdev**: “I would also incorporate alternate models which could help increase model performance.”
* **Ian Jackson**: “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.”
* **Jemin Vagadia**: “…but it would benefit from a more detailed discussion on… whether other models, such as random forests or gradient boosting, were considered for comparison.”

Several reviewers suggested incorporating alternate models for better performance comparison. For example, Uditi Namdev mentioned incorporating alternate models to improve performance, Ian Jackson suggested comparing the decision tree with a multiple linear regression for better context, and Jemin Vagadia recommended discussing other models like random forests or gradient boosting.

However, based on guidance from the professor and TAs, the project focuses on developing and evaluating a single model in depth. Our exploration and optimization of the decision tree model, including forward feature selection, aligns with this directive. In fact, we received concerns from Professor Eran Mukamel about incorporating 3 different models at one point (linear regression, decision tree, and support vector machine), so we stuck with a decision tree for this project.

Two people mentioned the problem that decision trees are prone to overfitting.

* **Jemin Vagadia**: “Decision trees are known to be prone to overfitting, especially with complex datasets like the Ames Housing dataset. The study mentions the use of grid search for hyperparameter tuning to mitigate this issue, but it would benefit from a more detailed discussion on how overfitting was specifically addressed and whether other models, such as random forests or gradient boosting, were considered for comparison.”
* **Yanghong Wei**: “Models like decision trees are prone to overfitting, and even though the study used grid search for hyperparameter tuning, there is limited discussion on how overfitting was explicitly addressed. The study could incorporate techniques such as pruning or using ensemble methods like random forests to reduce overfitting.”

In response to these concerns, we have expanded our discussion in the relevant sections of the paper to explicitly address how our techniques help mitigate overfitting. Specifically, we have added sentences in the sections on K-Fold Cross-Validation and Optimizing Hyperparameters with Grid Search to highlight how these methods contribute to reducing overfitting.

In the K-Fold Cross-Validation section, we now emphasize that k-fold cross-validation provides a more reliable estimate of the model's performance on unseen data by averaging the performance across multiple folds, thus ensuring the model is not overly tuned to any single subset of the data. This helps improve the generalization ability of the model and mitigates overfitting.

In the Optimizing Hyperparameters with Grid Search section, we have detailed how grid search, combined with k-fold cross-validation, not only identifies the optimal hyperparameter values but also reducing the risk of overfitting by ensures that the chosen model parameters generalize across different subsets of the data.

Some concerns regarding the generalizability of the Ames Housing dataset were brought up.

* **Jemin Vegadia**: “The study focuses on the Ames Housing dataset, which is a well-known dataset in the machine learning community. However, it lacks discussion on how the findings and the developed model can be generalized to other housing markets or datasets. Including an external validation with a different dataset or discussing the potential limitations in applying this model to other real-world scenarios would strengthen the study.”
* **Yanghong Wei**: “The dataset is specific to Ames, IA, and findings may not generalize to other regions. The study should explicitly discuss the limitations of geographic specificity and suggest directions for validating the model with diverse datasets from different locations.”
* **Yanghong Wei**: “The author is well aware of the limitations of the project and makes suggestions to improve their model. It may be worthwhile to discuss, among all the measurements that could help, what method is the most effective way of improving the robustness of the model.

We appreciate the feedback on the generalizability of our study. While we've addressed this in the limitations section, we've enhanced our discussion by adding comments on potential future directions. Our analysis primarily focuses on the Ames Housing dataset because it’s much more comprehensive and richer than other datasets, but we are aware it may not fully represent other housing markets. We acknowledge the need for caution in generalizing findings beyond Ames, and we've included suggestions for future research, such as the reviewer’s suggestion of external validation with datasets from diverse locations. By discussing these considerations, we aim to ensure the relevance and potential avenues for anyone willing to extend our findings to broader contexts.

Ian Jackson wrote some important details about our hypothesis that need addressing too.

* **Ian Jackson**: ““...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.”

Ian Jackson’s points are legitimate concerns and we appreciate them bringing these points up. In light of the comments, we acknowledge that listing techniques like feature selection, cross-validation, and grid search within the hypothesis may not be appropriate. Therefore, we have opted to remove mention of these techniques from the hypothesis. Our focus is on exploring the impact of various factors on house prices rather than demonstrating the improvement of the model through specific techniques.

The revised hypothesis now includes examples to describe the variables we mentioned, such as location, size of rooms, and overall quality. This aims to provide clearer expectations regarding the factors expected to influence house prices. Regarding Ian's point about outlining the nature of the effect for each factor, we now mention that we expect a positive relationship between the mentioned factors and house prices.

Additionally, there is another reviewer's comment regarding adding more details or descriptions about these variables that is somewhat relevant to this point. We plan to address this next.

Below is the aforementioned point of concern regarding adding more details or descriptions about these variables.

* **Yanghong Wei**: “The definition of some factors could be explicitly explained in the report so that readers can understand each feature better. For example, what is included in overall quality?”

We are aware that providing explicit explanations for certain factors can enhance readers' understanding of the features used in our analysis. However, due to the extensive number of variables (79 in total) with detailed descriptions, directly incorporating all definitions into the paper might lead to clutter. To address this concern, we have provided information directing interested readers to the comprehensive data description file (data\_description.txt) included in the Ames Housing dataset. This file contains detailed descriptions of all variables, including the variables we mentioned that fall under the category of “overall quality” in our hypothesis. Additionally, we have included a link to the Kaggle dataset in the bibliography section, allowing readers to access the dataset for further exploration.

There was one comment about elaborating on important values.

* **Ian Jackson**: “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.”

We have made the following changes to address this concern:

1. We have provided a quantitative definition of the importance values metric, specifically mentioning that it is calculated based on the mean decrease in impurity (MDI) method commonly used in decision tree models.
2. We clarified that the importance values are calculated based on how much each feature contributes to reducing impurity across all the nodes in the decision tree during the training process.

Concerns about the clarity of some parts of the methodology were raised.

* **Ian Jackson**: “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.”
* **Jemin Vagadia**: “Some sections of the methodology, particularly around the feature selection process and grid search parameter tuning, could benefit from clearer explanations and visual aids to enhance understanding.”

The section describing Figure 1 (Grid Search Parameter Range) was something we were debating on removing because we believed it didn’t add much to the writeup. Given that it is causing confusion for some readers, we decided to remove the figure and simplify the whole section. Now, we only mention the grid search parameter range and mention that we believe it’s appropriate for GridSearchCV because it covers the range of values that GridSearchCV hovered around.

We’re unsure which sections Jemin found unclear because they didn’t provide details on what they found unclear. Upon asking a few other people, the only section that seemed confusing to people was just the Grid Search Parameter Range section. Therefore, we decided to focus only on modifying this section.

Next, Ian Jackson mentions the lack of a related studies and background section

* **Ian Jackson**: “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.”

In response to this feedback, we have added a new section titled “Related Works” to our paper. This section highlights previous research efforts that have tackled the prediction of housing prices using various datasets and methods. Specifically, we have discussed works that focus on feature engineering and exploratory data analysis using the Ames Housing dataset.

Some peer reviewers have made comments about the difficulty of reading the text on figures and more visuals or analysis.

* **Jemin Vagadia**: “While the study includes some visualizations, additional plots such as feature importance charts and partial dependence plots would help in understanding the influence of each predictor on house prices.”
* **Yanghong Wei**: “It is noticeable that there is a lot of information about the analysis results on the figures included in the project report. It will definitely help the readability of axis labels and figure titles if the font size is increased. Overleaf or LaTex scale the figures while not adjusting the font size on the figures.”
* **Ian Jackson**: “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.”

We’ve regenerated all the figures and increased the font size as large as possible without obscuring the other text on the figure nor extending out of bounds of the figure size. Additionally, while we do agree that feature importance charts and partial dependence plots would help people understand the influence of each predictor on house prices, we chose to leave those out for this write up because of strange behavior in the updated results of our study. Specifically, we have left a comment about this in the future directions section of the writeup. Ian brought up great points of investigation, so we described in detail the approach he mentioned as a future direction. Moreover, Professor Eran Mukamel discussed that interpretation is not the focus of the project. In fact, he has told us that we didn’t need to include it in the first place, as the project is about model selection instead.

One minor comment was brought up about the units of MSE.

* **Ian Jackson**: “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.”

There’s a new sentence after the first time MSE was mentioned explaining that the units are in dollars.

There was a minor suggestion about maintaining the model.

* **Yanghong Wei**: “Except for working with collected data, it may also be a concern that how the model and data set can be updated or retrained periodically as response to the dynamic changings of the society.”

A paragraph dedicated to addressing this was added in the Research Limitations and Future Directions section.

Lastly, one reviewer suggested making the captions more descriptive for the figures and tables.

* **Yanghong Wei**: “The figures and tables are informative, but adding more detailed captions below them could enhance understanding. Providing explanations about what the results in each figure represent or what data is being presented can help the reader better grasp the context. Rather than simply labeling the figures, more descriptive captions would improve clarity.”

All the tables and figures have been updated accordingly to include more details.