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.