Airbnb Price Prediction Task

1. What machine learning modeling approach would be best suited for this task and why?

For this task, a **gradient boosting algorithm** such as **CatBoost** or **XGBoost** would be best suited. These algorithms are highly effective for structured/tabular data and excel at capturing complex relationships between features. Additionally, CatBoost handles categorical variables natively, reducing the need for extensive preprocessing. This is particularly useful for this dataset, which likely includes many categorical features.  
  
I went ahead and developed and tested both models on Train, Test and OOT as well as 5 fold CV.   
  
The following were the results.   
  
A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated  
  
Based on the CV Mean scores, I ultimately opted for the Catboost model.

2. How can we most effectively utilize the entirety of the dataset?

a. Are there any potential issues that may prevent us from using all the data?

To effectively utilize the entire dataset, we can:

* **Impute missing values**: Missing data can be imputed using median values for numerical features and 'Unknown' for categorical features.
* **Feature engineering**: Enhance the dataset by deriving new features such as time deltas from date fields or interaction terms between features.
* **Data normalization/standardization**: Scale features to ensure that algorithms perform optimally.
* **Extract key words**: Features such as ‘summary’, ‘space’, ‘description’, ‘neighborhood\_overview’, etc. may contain descriptions that correlate with the quality of the listings. Mining key words could result in more predictive features.

Potential issues that may arise include:

* **High cardinality of categorical features**: Some categorical features may have too many unique values, making them difficult to handle.
* **Long skewness**: If the target variable is imbalanced, this might skew the model's performance.
* **Multicollinearity**: Highly correlated features could affect the model's performance.

3. How do you plan to measure the success of the model?

The success of the model will be measured using:

* **Root Mean Squared Error (RMSE)**: This metric gives a sense of how far the predictions are from the actual values, penalizing larger errors more heavily.
* **R^2 Score**: This metric measures the proportion of variance in the dependent variable that is predictable from the independent variables. This is particularly helpful as it has an upper and lower bound.

Additionally, performance on train, test and OOT will be measured. The objective is to make sure that performances on train and test are similar. This would indicate that the model is not over/under fitted. The OOT (though a small number of observations), will show the model performance on data it has never been exposed to in train or test.

4. What should be the success threshold for deploying this model on the Airbnb website and why?

Business rationale:   
A reasonable success metric ultimately depends on how the model is used by end consumers and the system it replaces.

* What is the cost of having analysts manually assess prices?
* Does the improved accuracy of manual review (if any) justify the expense incurred to generate them?

Science based rationale:  
Though a simplistic measure of an acceptable R^2 is 0.7 on test sets and RMSE within 10-20% in test sets, these numbers can dramatically change based on the industry standard, domain, task and data quality.

5. Please provide all analytical code and supporting graphs for your approach.

The entire analytical process, including code for data preprocessing, model training, evaluation, and prediction, is documented in the following notebooks:

* **02\_preprocess**: Data preprocessing steps, including feature engineering and missing value imputation.
* **03\_eda**: Exploratory Data Analysis (EDA), with supporting graphs.
* **04\_model\_dev**: Model development and training code, including parameter tuning.
* **05\_inference**: Code for scoring new data using the trained model.

These notebooks are stored in their respective directories and contain all necessary code and visualizations.

6. If you were able to conduct any ML modeling, kindly share the code and results (this is optional).

The ML modeling was conducted using CatBoost, with the final model's code and results provided in the **04\_model\_dev** and **05\_inference** notebooks.

7. How much time did you spend on this assignment?

To meet the minimum requirements of this task, I spent just ~3 hours.

This includes:

* Preprocossing data
* Exploratory data analysis
* Fitting XG Boost model on the inference data

To go above and beyond, I spent ~6 hours.

This includes:

* Templatizing code and cleaning files
* Annotationg and commenting of the notebooks and functions
* Setting up virtual environments with compatible libraries for best practice
* Developing a second Catboost Model to further improve on the performance of XGB
* Cross validation and hyper parameter tuning
* Extensive validation process on train, test and OOT
* Rerun notebooks to ensure reproducibility
* Writing a succinct README