|  |
| --- |
| Project Name: |
| Real Estate Machine Learning Pricing Model |
| Seller-focused Technology |

|  |
| --- |
| 3-16-2024 |

**Problem Statement**

Buying a house is a complicated process. You need to shop around, compare multiple houses, and file countless documents for the loan. However, one aspect that people often overlook is: how justified is the price of this house? You could buy a mansion in Texas for $300k, but that same price will only get you a bathroom-sized apartment in New York City. So, the question that arises is whether this price is truly justified for this specific house? It’s not cheap to buy a house, so every penny needs to make its worth.

Similar complications arise on the seller’s perspective, where they are conflicted between getting the fair price for their home while having enough buyer interest. They don’t want to sell the house for too little, want to maximize their earnings, but not so much that they don’t get any buyers. Therefore, our team wants to use machine learning to build a model that addresses such questions and helps sellers identify the optimal price point, when selling their lovely home. To build our model, we will consider all the features associated with a house, such as:

* *Location of the neighborhood*
  + *Location can be derived from state level, zip code level, street level….*
* *Building Material/Style*
  + *House style (1 story? 2 story?), exterior, building foundation type (wood? concrete?) ….*
* *Size*
  + *Area within the house, total area of the property, numbers of bathrooms….*

There are many more features (~80 features given the dataset that we have chosen) in our dataset, which will be used to see their impact on the pricing. We will begin our initial modeling using a linear regression model and then adjust it as required based on its performance. Ideally, the final model/product can take an address from the user, prompt a price, and provide an explanation and relevant statistical reasoning behind the price. The model will help the seller to make more money by avoiding non-optimal pricing and save his time spent on conducting research to find this optimal price. All the house owners should also know their worth!

**Available Solution**

Currently, there are already people who utilize machine learning techniques to benchmark/price their property. For example, in the resource that we found, *Real Estate Pricing Prediction Using Machine Learning,* ([Dhandha, Pandey, et al., 2023](https://link.springer.com/chapter/10.1007/978-981-99-2710-4_9))the authors used simple linear regression model to reduce the risk of bias by controlling the optimal hyperparameter, and then used k-fold cross-validation on the cleaned dataset to improve performance. The objective of the study in the resource is exactly same as our proposed business model which is to build a model that can predict the price given the feature.

One advantage of using a technique like simple linear regression is its accessibility and ease of implementation. It offers a relatively straightforward approach to analyzing and predicting property prices, making them accessible to a wide range of users, including those without extensive technical expertise. Additionally, by employing techniques like controlling optimal hyperparameters and k-fold cross-validation, we can enhance the accuracy and reliability of our predictions, thereby providing valuable insights for sellers.

Despite these advantages though, using simple linear regression for real estate pricing prediction may have limitations. One significant drawback is its reliance on linear relationships between input variables and the target variable (property price). Real-world data often exhibits complex, nonlinear relationships that may not be adequately captured by linear models, potentially leading to inaccurate predictions. Furthermore, while techniques like k-fold cross-validation help mitigate overfitting and improve model performance, they may not completely eliminate the risk of bias or generalize well to unseen data. Therefore, we will have to exercise caution while building our model and experiment with different kinds of models to enhance performance.

It is also important to mention that one of the reasons why we picked this business plan is that beside the resource that we referenced here in the proposal, we think there are a lot more references that exists since real estate is a relatively mature business sector.

**Data Set**

To begin our initial analysis, we obtained our first dataset from the Kaggle and the dataset is split into a training and testing set. There is also a document of data description that explain all 81 columns that is in the data set. In total the file contains approximately 2000 rows of data. Even though it would be more ideal that we have a bigger dataset, but one thing that our team value a lot is data quality. The completeness of data is very high in the dataset, since it’s a very comprehensive set of features that could be used for detailed analysis for the predicting pricing model. For the detail of the data, you may reference the *data\_description* fIle.