A Report on the Efficacy of Multiple Artificial Intelligence Models to Predict Housing Prices

Project Goals

The objective of this project is to build, optimize and test multiple models of artificial intelligence for the purpose of estimating housing prices to within the nearest million dollars based on statistics such as square footage, main road access, number of bedrooms and bathrooms, etc. The parties which may be interested in the results of this project are those who wish to create state-of-the-art models for this defined task. Though the project defined here is not attempting to create such a model, the results of this project should indicate which models might be most promising for the performance of this task. If this project is successful, it may assist in the process of selecting which models to use for this task in the future.

Data Preparation

The process of data collection and preparation was the same for each artificial intelligence model to ensure that the models were being compared in equivalent environments in order to allow for accurate and useful comparisons between the metrics of each model. To begin, raw data was imported into a dataframe:

```
This is the raw data for the project:

        price  area  bedrooms  bathrooms  stories mainroad guestroom basement hotwaterheating airconditioning  parking prefarea furnishingstatus
0    13300000  7420         4          2        3      yes        no       no              no             yes        2      yes        furnished
1    12250000  8960         4          4        4      yes        no       no              no             yes        3       no        furnished
2    12250000  9960         3          2        2      yes        no      yes              no              no        2      yes   semi-furnished
3    12215000  7500         4          2        2      yes        no      yes              no             yes        3      yes        furnished
4    11410000  7420         4          1        2      yes       yes      yes              no             yes        2       no        furnished
..        ...   ...       ...        ...      ...      ...       ...      ...             ...             ...      ...      ...              ...
540   1820000  3000         2          1        1      yes        no      yes              no              no        2       no      unfurnished
541   1767150  2400         3          1        1       no        no       no              no              no        0       no   semi-furnished
542   1750000  3620         2          1        1      yes        no       no              no              no        0       no      unfurnished
543   1750000  2910         3          1        1       no        no       no              no              no        0       no        furnished
544   1750000  3850         3          1        2      yes        no       no              no              no        0       no      unfurnished
```

Next, the prices of the houses were rounded down to the nearest million dollars so that each unique label (the price) would have more accompanying data points than if the models had to predict the price to a much more accurate number of significant figures. This was the resulting dataframe:

This is the dataframe after prices were rounded to the nearest million dolalrs:

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	13	7420	4	2	3	yes	no	no	no	yes	2	yes	furnished
1	12	8960	4	4	4	yes	no	no	no	yes	3	no	furnished
2	12	9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished
3	12	7500	4	2	2	yes	no	yes	no	yes	3	yes	furnished
4	11	7420	4	1	2	yes	yes	yes	no	yes	2	no	furnished
..	...	...	...	...	...	...	...	...	...	...	...	...	...
540	1	3000	2	1	1	yes	no	yes	no	no	2	no	unfurnished
541	1	2400	3	1	1	no	no	no	no	no	0	no	semi-furnished
542	1	3620	2	1	1	yes	no	no	no	no	0	no	unfurnished
543	1	2910	3	1	1	no	no	no	no	no	0	no	furnished
544	1	3850	3	1	2	yes	no	no	no	no	0	no	unfurnished

Note that the first column now ranges from 1 to 13. This is a significant reduction in the number of labels. Finally, the dataset was label encoded and normalized. The label encoder was used to turn non-numerical answers into data which could be analyzed by an artificial intelligence model. Normalization is required for many of the models to be able to be accurately trained on the data. This was the resulting dataframe after these steps:

```
This is the resultant dataframe after all data preprocessing:

[[ 4.61400478  1.04672629  1.40341936 ...  1.51769249  1.80494113
  -1.40628573]
 [ 4.08516511  1.75700953  1.40341936 ...  2.67940935 -0.55403469
  -1.40628573]
 [ 4.08516511  2.21823241  0.04727831 ...  1.51769249  1.80494113
  -0.09166185]
 ...
 [-1.73207119 -0.70592066 -1.30886273 ... -0.80574124 -0.55403469
   1.22296203]
 [-1.73207119 -1.03338891  0.04727831 ... -0.80574124 -0.55403469
  -1.40628573]
 [-1.73207119 -0.5998394   0.04727831 ... -0.80574124 -0.55403469
   1.22296203]]
```

The data is now ready for use in our models.

Approach

The approach for this project is quite simple. For each desired model, set up the model, test the model, attempt to optimize the model, and record the best achieved metrics for each model. Though this approach is not complicated, the process of testing numerous models on the same data set does not require an overly complicated solution. One of the most likely anticipated problems in this approach is the difficulty in tuning each model for optimal performance. The number of variables which could be altered and the number of possible values for each variable means that there is a vast number of possible states for each model to be tested on.