Marylyn Chen

Shih-Huan Chou

Dongkyung Lee

Charline Truong

Finding the Perfect House: Utilizing Ensemble Approach

I. Problem

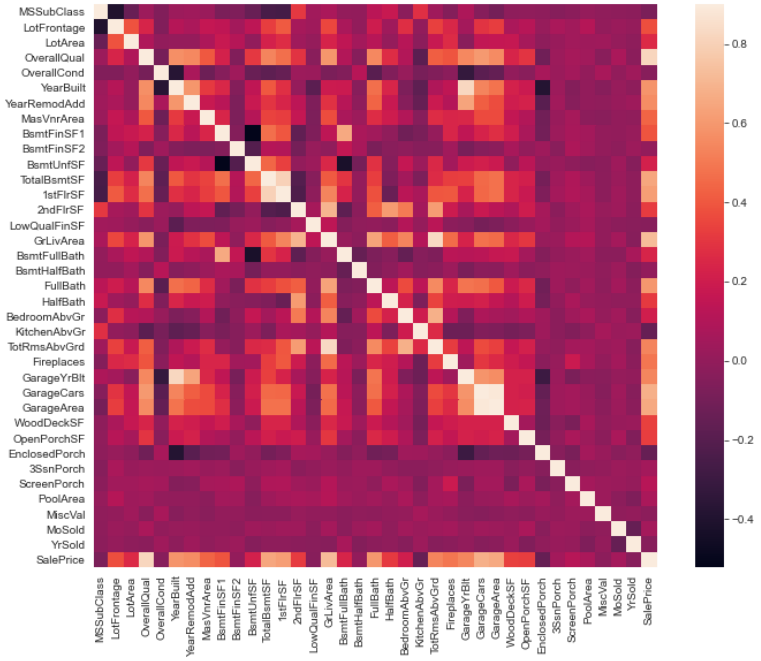
In today's economy, housing prices are constantly fluctuating. For potential buyers of new houses, it has become harder to determine the predicted price of their future home with their expected dream features. Factors such location, amenities, and current economic climate impact prices on the housing market usually impact the pricing of a house the most. However, many other factors such as the features of a home, number of rooms, number of bathrooms, and age of the house may also be very important in the eyes of a buyer. Therefore, accurately predicting the price of a house can help a buyer identify which undervalued houses are considered an investment.

II. Background

In this problem, we will be tackling the AMES dataset compiled by Dean De Cock. From this dataset, we are given a set of training data points that we can use to train our model and a set of testing data points without the sale prices. Our model will be trained from the training set and will be used to evaluate the testing data points and the accuracy of the model will be determined by the Root Mean Squared Logarithmic Error. Additionally, we will utilize the hedonic price model to forecast the predicted prices of the houses. In a hedonic price model, the value of a product is determined by features a product has, which in our case is the characteristics (garage space size, utilities, size of house, etc.) of a house (Calmasur, 2016, pg.255). Before approaching

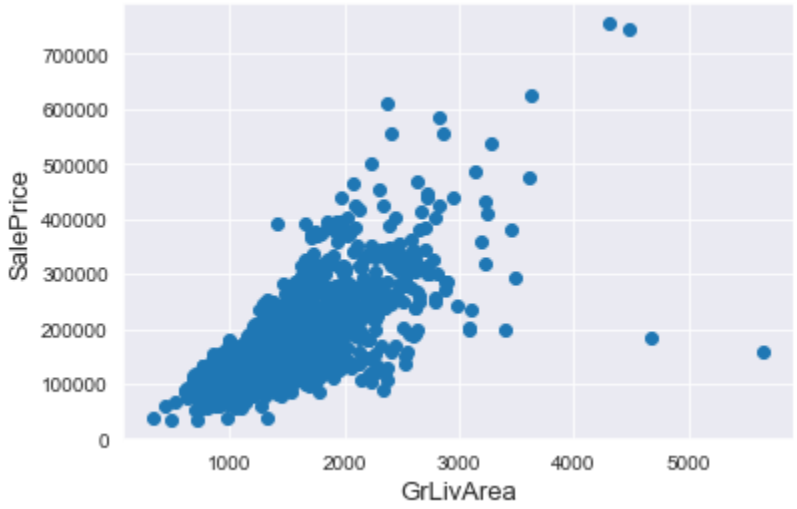
the problem head on, we must analyze the structure and patterns that the dataset exhibits in order to choose better models or to do appropriate feature engineering.

Correlation analysis is often the main approach for assessing promising features that relates well to the problem in order to reduce the complexity of the dataset by throwing out features that do not help much in categorizing or predicting outcomes. We can determine the correlation between different features by using the 'dataframe.corr()' method and determining which features to keep by setting a threshold. In the following figure, we graphed a heatmap of the correlation between different features. Lighter colors suggest more correlation between the two features. As you can see, a lot of the features do not actually correlate that well with the sale price and will just be another parameter that can cause overfitting.



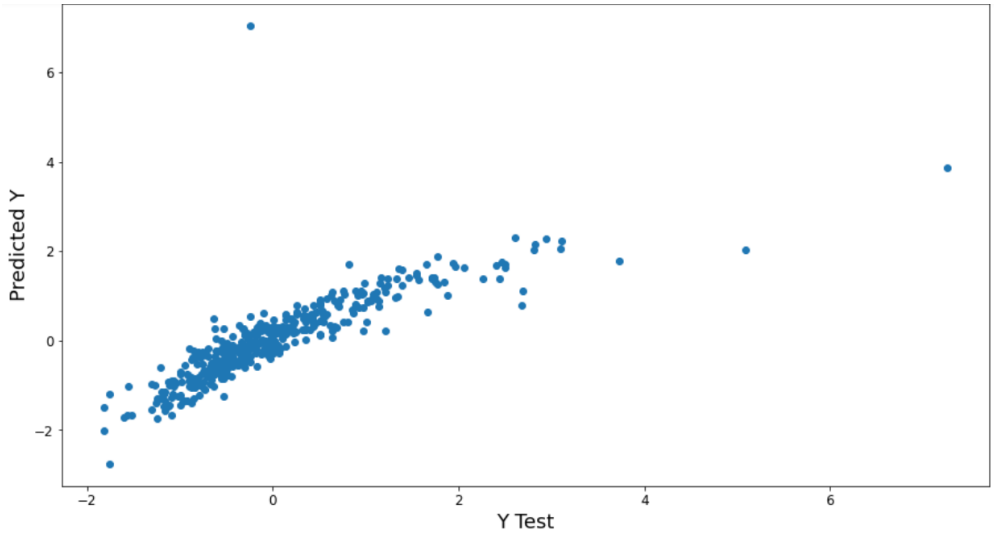
A data that correlates well with price can be evident when graphed, and will typically show a positive correlation when graphed with the Sales price. Graphing these correlations will also reveal unwanted outliers from the data that we will remove in order to increase the quality of the data and decrease the possibility of overfitting. A good example of a factor that correlates

well and has a well-behaved graph is the amount of living area as shown below. The figure below almost shows an almost linear relationship between the two and contains two outliers that we later removed. These outliers show large living areas but low prices, suggesting that these transactions do not represent the data well enough to be included and the inclusion of these data may be detrimental to our model.



III. Baseline Approach

For the baseline, we decided to use the linear regression model from the sklearn package in Python because it is one of the simplest approaches. Linear regression is useful for our purpose because we are using numerical features such as size, rating, age, etc. (categorical data is one-hot encoded into binary numbers) from houses to predict their prices. All columns were scaled so that no column is weighted more than another.



As we can see, our predicted values are approximately equivalent to our actual values from our test set. Our mean average error, mean square error, and root mean square error are as follows:

```
MAE: 0.2843756853965748
MSE: 0.29321173078045865
RMSE: 0.5414902868754514
```

IV. Ensemble Approach

While deciding on which approach would best serve our problem, we decided on the ensemble approach because it can create lower variance and lower bias. Each additional model can be thought of as like a data point, and more data points result in lower variance. Additionally, past research has shown that an ensemble will typically produce an accuracy greater than any of the single classifiers used in the ensemble (Maclin & Opitz, 1999, pg.169). The specific type of ensemble approach we will be utilizing is the stacked generalization or STACK. In STACK, we combine different prediction models in a single model, and the model will work at different levels (Coelho & Ribeiro, 2020, pg 6). We chose the STACK approach because we are able to include a diversity of models in order to improve our accuracy without being constrained to a

specific type of classifier. The base models that we decided to combine in our ensemble approach are Elastic Net Regression, Gradient Boosting Regression, Kernel Ridge Regression, and Lasso Regression. These models are from the sklearn package in Python, and the mean-squared error and the standard deviation are shown below in the images. These four models already give better mean-squared error on the test set compared to the base model (linear regression) and combining them should give us even better results.

```
Lasso score: 0.1115 (0.0074)
```

```
ElasticNet score: 0.1116 (0.0074)
```

```
Kernel Ridge score: 0.1153 (0.0076)
```

```
Gradient Boosting score: 0.1168 (0.0082)
```

We averaged the predicted prices of the four models and compared to the actual price. As expected, the averaged base models performed with better accuracy and lower variance than all the other base models as shown below. The averaged base models also have a lower chance of overfitting since we are not depending on just one model. Comparing this mean-squared error to the leaderboards in Kaggle, we would be somewhere around 106th place. Considering the implementation of the model, the results are surprisingly good.

```
Averaged base models score: 0.1092 (0.0077)
```

V. Conclusion

In recent years, the hedonic price model has come under scrutiny. This is because hedonic pricing models function under arbitrary assumptions like homogeneity of the housing product, and there are no interrelationships between the implicit prices of attributes. Both of these assumptions can be arguable because it would be more accurate to classify houses as heterogeneous because they can be differentiated by different categories, and the attributes—in our case housing features—will not give the same level of utility or disutility to all buyers (Chin & Chau, 2003, pg. 149). However, the disadvantages of the hedonic pricing model do not outweigh its merits. It's straightforward, and the personal characteristics of the buyers and sellers do not need to be held in account. Moreover, it's flexible and adaptable to the many factors and parameters that go into pricing a house (Chin & Chau, 2003, pg. 150). Although the ensemble approach has shown to perform better than linear regression, there are still a few drawbacks to utilizing the ensemble approach. These drawbacks include computational costs, difficulty in defining the models for the composition of a STACK approach, and difficulty in determining values for the set of hyperparameters (Coelho & Ribeiro, 2020, pg 6). While these drawbacks may seem daunting, the performance results of the ensemble approach justifies the increased complexity. As seen above, the mean-squared error of our ensemble approach to the test set is 0.1092 compared to our baseline approach of linear regression of 0.2932. {insert something about numbers}. Overall, while the ensemble approach does have its disadvantages, it allows for better accuracy, lower variance, and lower bias compared to other baseline models.

References

Calmasur, G. (2016). Determining factors affecting housing prices in Turkey with Hedonic pricing model. In *International Conference on Business and Economics Studies, Washington DC, USA* (pp. 255-269).

Chin, T. L. and Chau, K. W. (2003). A critical review of literature on the hedonic price model, International Journal for Housing and Its Applications 27 (2), 145-165.

Coelho, L., Ribeiro, M. (2020) Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. In *Applied Soft Computing. Vol 86,* 1-17, https://doi.org/10.1016/j.asoc.2019.105837

Maclin, R., Opitz, D. (1999) Popular Ensemble Methods: An Empirical Study. In *Journal of Artificial Intelligence Research. Vol 11,* 169-170, https://doi.org/10.1613/jair.614