**A Matter of Price:**

**Housing Data and Machine Learning**

Caleb Bush, [cbush6@bellarmine.edu](mailto:cbush6@bellarmine.edu)

Ben Gierhart, [wgierhart@bellarmine.edu](mailto:wgierhart@bellarmine.edu)

1. **INTRODUCTION**

This project uses the [Housing Prices Dataset](https://www.kaggle.com/datasets/yasserh/housing-prices-dataset) from Kaggle, which includes data from northeast states in the U.S.. The dataset has a number of variables in it such as price, area, bedrooms, bathrooms, stories, etc. This analysis is using furnishingstatus (furnished, semifurnished, or unfurnished) to predict the price of a home.

1. **BACKGROUND**

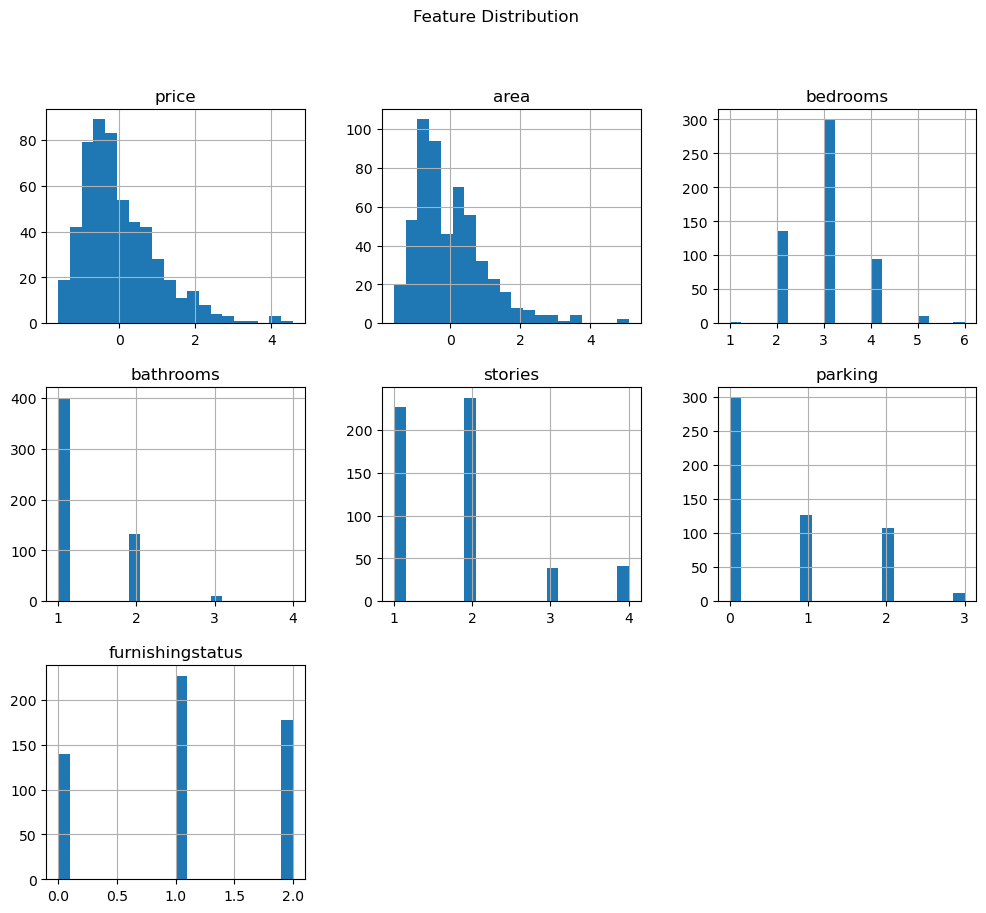
The purpose behind the creation of this dataset is to educate the analyst with opportunities to explore a small dataset with strong multicollinearity. It is reasonably well established that different variables for homes in the housing market can strongly influence price. This dataset allows the analyst to substantiate their intuition with data and machine learning.

1. **EXPLORATORY ANALYSIS**

This dataset contains 545 samples with 13 columns with various data types.

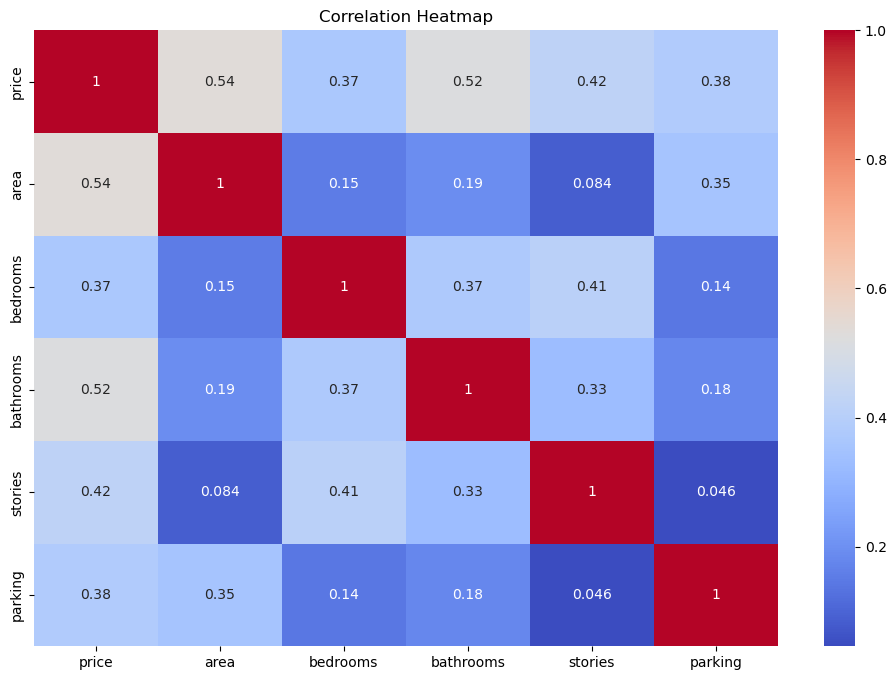
**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| price | integer |
| area | integer |
| bedrooms | integer |
| bathrooms | integer |
| stories | integer |
| mainroad | object |
| guestroom | object |
| basement | object |
| hotwaterheating | object |
| airconditioning | object |
| parking | integer |
| prearea | object |
| furnishingstatus | object |

**

**Figure 1:** Histograms of numerical data.

The price and area variables are right skewed**,** as are number of bathrooms and parking. The furnishing status distribution is approximately normal.

****

**Figure 2:** Heatmap of pre-encoded variables

**A screenshot of a graph

Description automatically generated**

**Figure 3:** Heat map of encoded variables.

Figure 2 shows a strong correlation between area and price and bathrooms and price. Figure 3 shows a strong correlation between area and price, bathrooms and price, and aircondonditioning\_yes and price.

1. **METHODS**

In this section, describe how you prepared the data for your model and performed multiple experiments using different parameters for the model(s).

* 1. *Data Preparation*

In this experiment, there was a n initial check for nulls. Since there were none, the next step was to use .describe() to look at counts, mean, std, min, max, etc. Since ‘furnishingstatus’ was the variable used to train the model, the variable was encoded so that furnished=0, semifurnished=1, and unfurnished=2. Also, since the ranges of the price and area variables were so much wider than those of the other variables, they were scaled. This scaling transforms the numerical data so that the mean is 0 and the standard deviation is 1. The previously large scale of the numbers was influencing our model and causing large MSE and MAE.

* 1. *Experimental Design*

Table 2: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | “Simple” Prediction Test using a sample row, all variables |
| 2 | 80/20 split with all variables for train, and test |
| 3 | 70/30 split with all variables for train, and test |
| 4 | Selected features with 80/20 split for train, and test |

* 1. *Tools Used*

The following tools were used for this analysis: JupyterLab 3.6.7 and Python v3.11.7 running the Anaconda 4.3.22 environment for Apple Macintosh computer was used for all analysis and implementation. In addition to base Python, the following libraries were also used: pandas 2.1.4, seaborn 0.13.2, matplotlib 3.8.0, numpy 1.26.3, and sklearn 1.2.2.

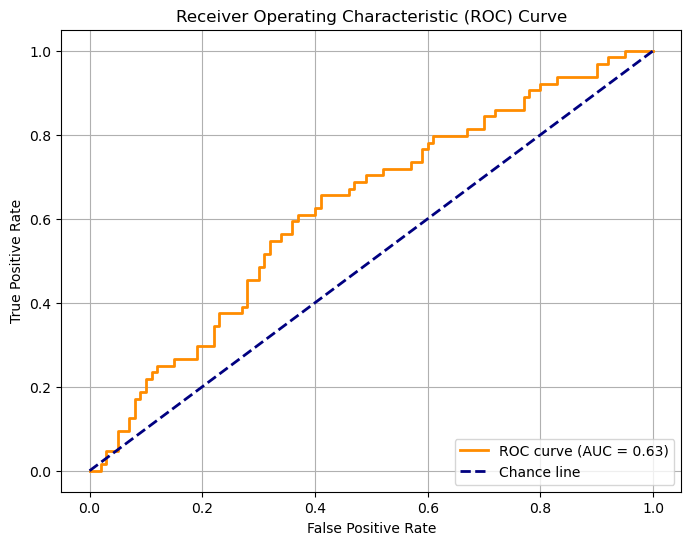
Pandas, seaborn, matplotlib, and numby were imported to create graphical representations of the data. Sklearn was used to train and test models, execute linear and logistical regressions, create confusion matrices, encode categorical data, create ROC curves, and scale data.

1. **RESULTS**
   1. *Classification Measures/ Accuracy measure*

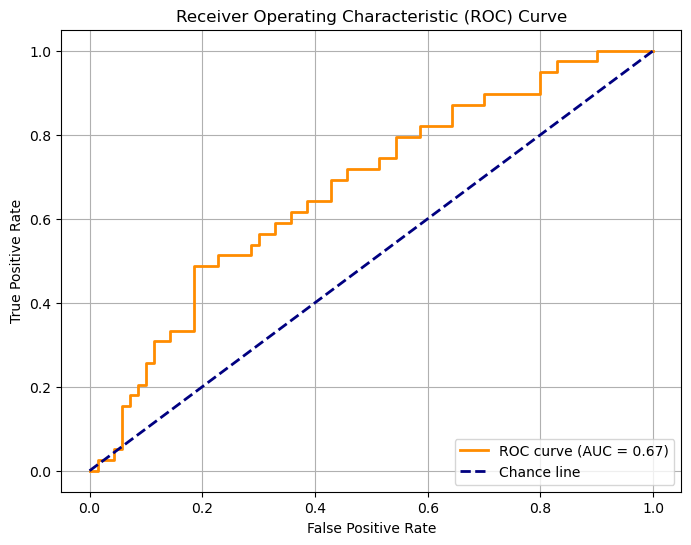
A graph with orange and blue lines

Description automatically generated

**Figure 4:** ROC Curve (80/20 split with all variables for train, and test)



**Figure 5:** ROC Curve (70/30 split with all variables for train, and test)



**Figure 6:** ROC Curve (Selected features with 80/20 split for train, and test)

* 1. *Discussion of Results*

The area under curve (AUC) was lowest (0.63) with the 70/30 split. This is because there was still sufficient data to train the model and more information to test than the 80/20 split. The AUC was highest with the 80/20 split. The hypothesis is that this is because there wasn’t as much data to test, and all features were used. The AUC for special selected features was slightly better (0.67).

* 1. *Problems Encountered*

No project goes perfectly smooth. Discuss any problems you had with obtaining the data, preparing the data, implementing the model, or evaluating the model. **It would be highly unusual to indicate that you had no problems.**

We encountered problems with needing to encode the ‘furnishingstatus’ feature as well as other categorical data. We also first experienced high MSE and MAE, so we scaled the ‘price’ and ‘area’ features so that their vast range stopped overinfluencing the model. We also experienced problems inversing the scaling when we wanted to check the accuracy of the model to the original data.

* 1. *Limitations of Implementation*

While the model shows moderate performance with an R² score of about 65%, the MSE and MAE values indicate there’s still significant room for improvement. A model with a better R² score would be better.

* 1. *Improvements/Future Work*

Lower MSE and MAE would be ideal for a more accurate model, so future efforts should focus on reducing these errors. Fine-tuning the model, improving feature selection, or exploring more advanced regression techniques could potentially help improve these metrics.

1. **CONCLUSION**

Our mission was to use the [Housing Prices Dataset](https://www.kaggle.com/datasets/yasserh/housing-prices-dataset) from Kaggle. The dataset has a number of variables in it such as price, area, bedrooms, bathrooms, stories, etc. This analysis is using furnishingstatus (furnished, semifurnished, or unfurnished) to predict the price of a home.

We gave ourselves a challenge by encoding the data of the categorical variables and scaling the data, but we ultimately succeeded in building a model that works. It’s not a bad model, but it could definitely be better by implementing the strategies laid out in section V.E.

**REFERENCES**

<https://www.kaggle.com/datasets/yasserh/housing-prices-dataset>