**Predicting the Housing Market with MLR**

Langdon Hatton [lhatton@bellarmine.edu](mailto:lhatton@bellarmine.edu)

Ashley Ridley [aridley@bellarmine.edu](mailto:aridley@bellarmine.edu)

**ABSTRACT**

In this project, we performed multiple linear regression on a housing dataset using R and python. The dataset we chose consists of 545 entries with 13 total columns and contains various types of variables such as categorical. We will use this dataset to test a prediction based on the independent and dependent variables we choose. We will also find the intercept, coefficient, R-square and RMSE and compare the prediction to the actual validation to see if our model is accurate. Included in our project will be visual representation of our data to show any interesting findings, while displaying any outliers.

1. **INTRODUCTION**

For this team project, we will be using multiple linear regression to predict the price of houses from our dataset. We will be using a variety of techniques to predict the price of the houses in our dataset. We will also be changing our test cases to see if we can make our models more accurate.

1. **BACKGROUND**
   1. *Data Set Description*

Because we wanted to use MLR, we found a dataset that had a value that would make sense to predict. We found a housing dataset that had multiple varieties of houses, such as the number of beds/baths, area, furnished, etc. The value that we chose to predict was the price of the house which would be based on independent variables. This dataset has no missing values, and it looked balanced. This makes it a good choice for our project.

* 1. *Multiple Linear Regression*

A statistical model known as multiple linear regression (MLR) is used to determine the relationship between a dependent variable and several independent variables. Predicting the value of a dependent variable using the values of two or more independent variables is the objective of this model. MLR makes the supposition that there is a linear relationship between the independent and dependent variables. The relationship between the variables is represented by a linear equation that the model generates, with the coefficients denoting the relationship's strength and direction.

Finding the best fit line that correctly predicts the value of the dependent variable based on the values of the independent variables is the fundamental tenet of MLR. The coefficients of the independent variables in the linear equation are estimated by the model using the least squares method. For each data point, this entails minimizing the sum of the squared differences between the dependent variable's expected and actual values. The model presupposes that the independent variables and dependent variables have an additive relationship, and that each independent variable's effect is independent of the others. MLR can be applied in the real world to make predictions or determine how the variables are related. It is common practice to analyze data and make predictions based on that data in disciplines like finance, economics, and social sciences.

1. **EXPLORATORY ANALYSIS**

Our dataset has 545 entries. It contains 13 columns. There are no missing values in our dataset.

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data type* |
| *price* | *int64* |
| *area* | *int64* |
| *bedrooms* | *int64* |
| *bathrooms* | *int64* |
| *stories* | *int64* |
| *mainroad* | *object* |
| *guestroom* | *object* |
| *basement* | *object* |
| *hotwaterheating* | *object* |
| *airconditioning* | *object* |
| *parking* | *int64* |
| *prefarea* | *object* |
| *furnishingstatus* | *object* |

**Table 2: correlation matrix for the raw dataset**

price area bedrooms bathrooms stories parking  
price 1.000000 0.535997 0.366494 0.517545 0.420712 0.384394  
area 0.535997 1.000000 0.151858 0.193820 0.083996 0.352980  
bedrooms 0.366494 0.151858 1.000000 0.373930 0.408564 0.139270  
bathrooms 0.517545 0.193820 0.373930 1.000000 0.326165 0.177496  
stories 0.420712 0.083996 0.408564 0.326165 1.000000 0.045547  
parking 0.384394 0.352980 0.139270 0.177496 0.045547 1.000000

**Table 3: correlation matrix for dataset with prefarea removed**

price area bedrooms bathrooms stories parking  
price 1.000000 0.491259 0.361712 0.517225 0.483447 0.385977  
area 0.491259 1.000000 0.145326 0.196560 0.119140 0.304094  
bedrooms 0.361712 0.145326 1.000000 0.361541 0.476191 0.129084  
bathrooms 0.517225 0.196560 0.361541 1.000000 0.349467 0.224008  
stories 0.483447 0.119140 0.476191 0.349467 1.000000 0.095062  
parking 0.385977 0.304094 0.129084 0.224008 0.095062 1.000000

1. **METHODS**

To prepare this data we first had to find any missing values. We used this method first because missing data can cause issuses in the accuracy of the prediction model. Next, we used a correlation and heatmap to identify how well, or not so well, features within a dataset correlate with each other and whether it is positive or negative. After this visual representation of the data, we were able to split the dataset into independent and dependent variables where we could work with the categorical data and later turn the categorical columns into machine code (1,0) with the get\_dummies() function.

* 1. *Data Preparation*

There were no missing values in our dataset. The data was also balanced. All we needed to do to prepare our dataset was split the independent and dependent variables up, so we could test.

* 1. *Experimental Design*

Table X: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | Done in python with the raw dataset. This experiment used an 80/10/10 split |
| 2 | Done in python with normalized data. The prefarea column was removed. 80/1010 split |
| 3 | Done in Rwith the raw dataset. This experiment used an 80/10/10 split |
| 4 | Done in python with normalized data. The prefarea column was removed. 80/10/10 split |

* 1. *Tools Used*

Describe all of the software tools you used to perform your data preparation and model implementation. For example:

The following tools were used for this analysis: R running in R studio environment: Tidyverse which is an R programming package that helps to transform and better present the data, catools which converts the R vectors while encrypting binary data, ggplot which is used for the visualization of the data such as bar plots, Python running in Anaconda environment with Pandas which brings the data analysis to the current environment, matplotlib brings data visualization and graphical plotting to python, seaborn provides a high-level interface for drawing statistical graphics, sciketlearn is a simple and predictive tools for data analysis.

1. **RESULTS**
   1. *Mean square Error and R-Square calculation*

Experiment 1 --> The r-square is 0.72  
The RMSE is 994996.85

Experiment 2--> The r-square is 0.61  
The RMSE is 1173809.36

Experiment 3 --> R-squared: 0.6818

Mean squared error 1111187722284.4

Experiment 4--> R-squared: 0.6628

Mean squared error 1177569065986.21

* 1. *Discussion of Results*

Discuss which of your models provided the best classification (or some other outcome if not classification). Explain why you think your best model was the best and why your worst model was the worst.

We think that experiment 1 with an 80/10/10 split was the best model because the r-sqaure is the highest meaning that the model fit our dataset the best. The worst for our model was experiment 2 where the prefarea column was removed causing the columns to shift from 13 to 12 which we think made the data unbalanced for the split.

* 1. *Problems Encountered*

It was a bit difficult to find a dataset with good data. We eventually found one, but it definitely would have been better if it contained more entries. It was also very hard to use any visual representation in R. ggplot was not working correctly.

* 1. *Limitations of Implementation*

We do think we picked the best model to use for our experiments. MLR is the best model that we have learned for predicting a variable that is continuous. I do think our dataset was a bit small. Other than that, we did the best with what we know. We did run into some problems visualizing the data in R.

* 1. *Improvements/Future Work*

Starting with the dataset, we can try to find a better set. Next, we could change the data based on correlation and remove variables that do not correspond with the price of the house as much. We also could experiment with other regression techniques.

1. **CONCLUSION**

With the data set that we found, overall, it was a very good set for our models. The data was balanced making it a lot easier to handle the data, especially with the categorical variables. Unfortunately, while using R there were setbacks using bar plots, scatter plots and heatmaps to better display the data, but python was a better choice for using the visual aspect of the graphs. Our prediction vs actual was close in numbers, especially since our r-squared values were closer to 1, meaning that the models we used were a better fit for our models.

Overall, our models were okay at best. Experiment had the highest R square value at 0.72. Even this value is not super high. In the future our choice of dataset and regression techniques will hopefully help us improve our models.