House Price Prediction

Student Name

Name of Institution

House Price Prediction

Regression modelling was used to predict house prices using a sample of 300 observations, split into training and testing datasets. A pseudo-random sample of 50 houses was selected from the overall dataset using the RANDBETWEEN function. The 50 houses formed the testing dataset with the remaining 250 observations used to train the regression model. An initial regression model was fit to the training data using all 10 predictors. The model was highly significant (p < 0.001) explaining an adjusted 74.9% of the variation in house prices. However, not all the predictors had a significant effect on house price after accounting for the effect of competing predictors. Only house size (in m^2), bathrooms, parking, and yard were significant predictors of house price at a 5% level.

The non-significant variables were dropped from the regression model sequentially, beginning with Terrase, which was the most non-significant (p = 0.81). At each stage, the significance of each predictor was checked and the most non-significant predictor was dropped. After 5 reruns, the analyst arrived at the final model whose predictors – house size, atico, parking, and yard – were all significant at a 5% level. The combination of variables explained an adjusted 74.89% of the variation in the house prices. The predictive value of the regression equation, given as: house price = -56,827.77 + 3,443.77 \* size + 28,657.34 \* atico + 37,473.95 \* parking + 48,594.74 \* yard, was tested in the testing dataset. The average prediction error was $58,897.15, equivalent to 25% of the average house price in the testing dataset.

Number of bathrooms was a significant predictor of house price from the initial to the semi-final model, and only lost significance after the removal of number of rooms from the model. The regression analysis was, therefore, repeated using house size, parking, yard, and bathrooms. Bathrooms was non-significant (p = 0.051) at the selected alpha level of 0.05; thus, the model was not fitted to the testing dataset. Instead, the analyst developed three alternative models using subsets by sequentially dropping the predictor with the highest p-value from the final model. Each of the alternative regression models was fitted to the testing data and used in house size prediction. The predictive value of the alternative models was compared to that of the final model using the average error statistic, presented in the table below. The alternative model with three predictors – house size, parking, and yard – resulted in the smallest average error value. It was, therefore selected as the optimal model and the corresponding regression equation, house price = -54,842 + 3,457.48 \* size + 45,945.99 \* yard + 35,781.9 \* parking, was used to predict house prices in the provided test set.

|  |  |  |
| --- | --- | --- |
| Model | Predictors | Average error |
| Final model | House size, atico, parking, and yard | 58,897.15 |
| Alternative 1 | House size, parking, and yard | 57,878.45 |
| Alternative 2 | House size and yard | 59,607.15 |
| Alternative 3 | House size | 59,852.97 |