Lareisha Lewis

EDA:

In this explanatory analysis I will explore the home price dataset to determine which variables have the most impact on whether a home is priced greater than $250,000. This dataset consists of 13 variables that give information on 867 different homes. I built a model with the variable “250\_plus” as the response variable, and the other twelve variables as independent variables. The “250\_plus” variable has the values, “yes” and “no” to classify homes priced greater than $250,000 and homes that are not. The output of this model produced an AIC of 393.88. Based on the p-values some of the variables are insignificant. I proceeded to build the next model with the following significant independent variables; “GarageCars”, “Fireplaces”, “Baths”, “SqFeet”, “Age”, “LotSize”, “HouseStyle”, and “condition”. This model showed slight improvement as it produced an AIC of 392.74. While the 1.14 difference shows improvement, it is not a very drastic improvement. I continued to examine the p-values for verification and noticed only one value from the “HouseStyle” variable proved to be significant. To decide which values I should filter out, I built a third model with “250\_plus” as the dependent variable and “HouseStyle” as the independent variable. The “HouseStyle” variable states whether the home is 1 story, 2 story, etc. This model showed me that out of the seven categories, only two are significant. Since “1Story” and “2Story” were the only significant categories I filtered the data to represent only these two home styles. I also felt it was important to use these categories because they make up most of the dataset. The bar graph below displays a count for each category, and the portion of homes above $250,000 for each category. As you can see majority of the dataset is made up of 1 and 2 story homes under $250,000.

Chart, bar chart

Description automatically generated

I used the filtered dataset to build my fourth model with the same independent and dependent variables as the second model. This model produced all significant variables, an AIC of 360.42, and the residual deviance decreased. I concluded this model is the best fitted model based on the results.

To further validate my model, I built a ROC plot. The graph below are the results of the ROC plot. The plot indicates a strong model with the area under the curve sitting at 0.9558. The plot visualizes an increase in sensitivity, while specificity decreases. In other words, the true positive rate is increasing while the false positive rate is decreasing. To get the sensitivity, specificity, and accuracy of the model, I made predictions for the probability of default on all home styles except 1 and 2 story homes and built a confusion matrix. The confusion matrix gave me a true positive value of 61, true negative value of 24, a false positive value of 152, and a false negative value of 630. Using the TP, FP, TN, FN values with a 30% prediction cut-off, I was able to calculate 631 as the sensitivity of the model, and 153 as the specificity.

Chart

Description automatically generated

I went on to explore the parameter estimates of each variable to determine the odds of each variable impacting the response variable. I noticed many of the variables had low odds sitting around 1, however a few stood out above the others. The number of baths in a home, the style of a home, and the age of a home contained odds less than 1. “Baths” produced a log odds ratio of 0.27, “HomeStyle” produced 0.30, and “Age” produced 0.89. This indicates that as these variables increase, they are not likely to be good predictors for whether a home will be priced above $250,000. The number of fireplaces a home has, and the condition of a home on a scale between 1 to 10, both contain odds higher than all the others. “Condition” has an odd of 2.28, and “Fireplaces” has an odd of 3.95. This means that as the condition of a home increases by one unit, the odds of the house being more than $250,000 increases by 2.28. If the number of fireplaces a house contains increases by one unit, the odds of the house being more than $250,000 increases by 3.95.

Conclusion:

Based on the maximum likelihood the square foot of a home is the most significant variable to determine the probability of default. It contains a p-value less than 0.05, and its positive parameter estimates indicate an increase in the probability of a home being over $250,000 as the square feet increase. The age of the home is another significant variable, but its negative parameter estimates states that as the age of the home increases the home is less likely to be priced above $250,000. The plots below show the logistic regression model fitted for some of the variables from the model.

|  |  |
| --- | --- |
| Chart, line chart  Description automatically generated | Chart, line chart  Description automatically generated |
| Chart, line chart  Description automatically generated | Chart, line chart  Description automatically generated |

Text, table

Description automatically generated

Code:

Final Model

* train\_data <- home\_data %>%

filter(HouseStyle == '1Story' | HouseStyle == '2Story')

model2 <- glm(`250\_plus`~ GarageCars + Fireplaces + SqFeet + Age + condition + Baths + HouseStyle +LotSize, data = train\_data, family = "binomial")

Odds Ratio

* round(exp(coef(model2)), 5)

Predictions/ Confusion Matrix

* test\_set <- home\_data %>%

filter(HouseStyle != "1Story" | HouseStyle != "2Story")

predictions <- predict(model2, newdata = test\_set, type = "response")

* glm\_pred1 = rep("YES",867)

glm\_pred1[predictions>0.20] = "NO"

table(test\_set$`250\_plus`, glm\_pred1)

A picture containing graphical user interface

Description automatically generated

ROC curve

* prob = predict(model2, type = "response")

train\_data$prob = prob

roccurve = roc(`250\_plus`~prob, data = train\_data)

plot(roccurve)

Sqft vs. 250\_plus Logistic Regression plot

* ggplot(df, aes(SqFeet, `250\_plus`)) +

geom\_point(alpha = .15) +

stat\_smooth(method="glm", se=FALSE, method.args = list(family=binomial),

col="red") +

ggtitle("SqFt vs. $250,000 +") +

ylab("Probability of Default")+

theme\_bw()