MSCI-718 INDIVIDUAL ASSIGNMENT 4

PROBLEM STATEMENT AND DATA USED

Australia's real-estate has been hot right now and since getting into real estate has been on my mind since always, I have devised some strategies to understand the factors that affect the market first. The first thing that comes to my mind about real estate is the location, but the question is, is that all that influences the price? Hence, using the Melbourne housing dataset I am determined to discover three factors (including location) that characterize the housing price.

PLANNING

After importing the dataset, the three main factors that I feel reflects any house are their location, number of bedrooms and the type of the house. Hence the main variables are 'Price' (output variable), 'Suburb', 'Bedroom2 and 'Type' (predictor variable). I planned on comparing two models, one with only location and the no. of bedrooms and the other with an added variable that is type of house.

The data was wrangled and checked for normality assumption, which showed a positive skewness for data under 'Price'. Hence after performing logarithmic transformation, the data was now tidy and ready to be used in the model.

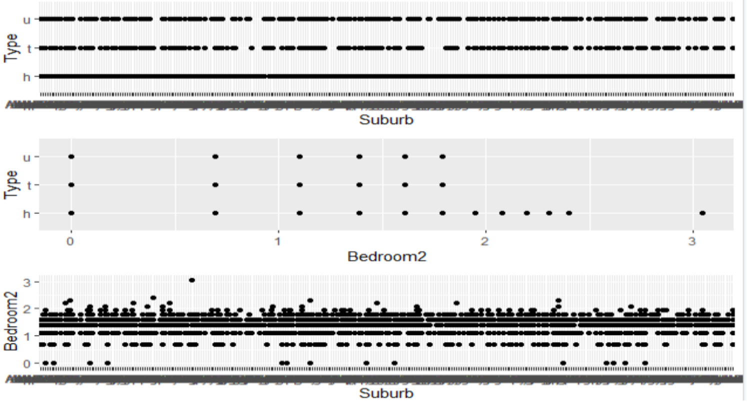

Figure: 1 – Graph for Multicollinearity checks

Before constructing the model, I verified the test for multicollinearity, since there are multiple input variables. From the figure:1, a visual inspection of the scatterplot of predictor variable suggests there is no correlation between them. Hence, I continued my analysis under the assumption of no multicollinearity that there are no external variables under consideration.

ANALYSIS

I plan to conduct multiple linear regression on the two model to find the best fit model and factors for predicting or analyzing the housing price market in Melbourne. After the model is developed, I will compare and conclude the best model followed by checking for assumption like

- Homoscedasticity of residuals (constant variance),
- Independence of residuals (test with Durbin-Watson),
- Linearity
- Outliers
- Influencers

Using the lm() function, I regressed for the formula "Price ~ Suburb + Bedroom2"(Model 1) and "Price ~ Suburb + Type + Bedroom2" (Model 2). From the summary of the regressed models

MODEL 1: The intercept, Suburb and Bedroom2 are significantly different than zero(p = $2.2e^{-16}$). The R^2 value of 0.6478 and an adjusted R^2 value of 0.6395 which are almost close to each other and suggests a 64.78% variance.
MODEL 2: The intercept, Suburb, Bedroom2 and Type are significantly different than zero(p = $2.2e^{-16}$). The R^2 value of 0.7752 and an adjusted R^2 value of 0.7698 which are almost close to each other and suggests a 77.52% variance.

To check if these two models satisfy the assumptions, to check if the residuals are homoscedastic and linear in nature, I plotted the residuals vs fitted value graph as below:

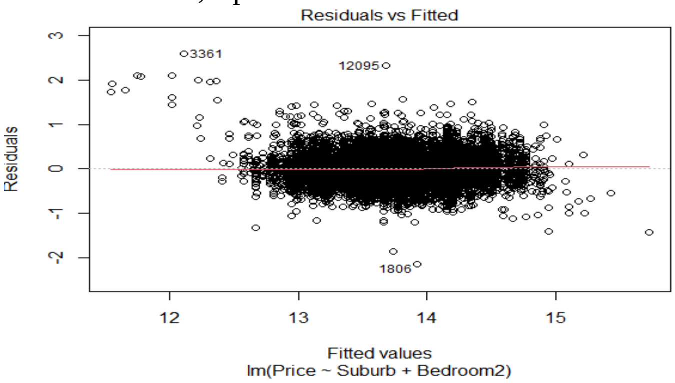

Figure 2: MODEL 1 Residual vs Fitted values

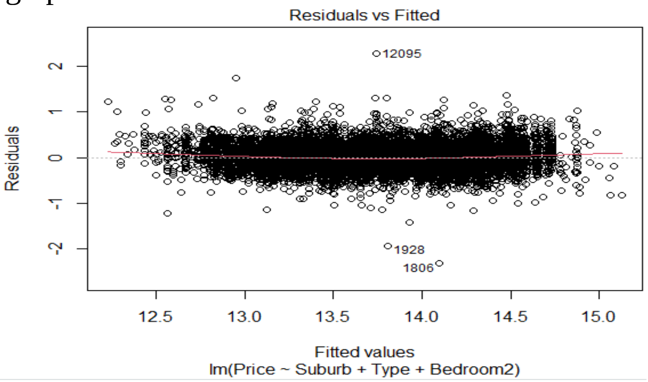

Figure 3: MODEL 2 Residual vs Fitted values

The above two figure 2 and figure 3 clearly suggest that the multiple linear regression proves that the residuals are homoscedastic and follow a linearity since the scatter plot is spread evenly on both sides on the center line.

I continued with the Durbin-Watson test to check the independence of residuals, which was not significant at the 5% level of significance (Model 1: d=1.970544, p=0.002) and (Model 2: d=1.982567, p=0.018). As d is very close to 2, we do not reject the null hypothesis and continue with the assumption of independence met.

To check for possible outliers, I tried to find the possible number of standardized residuals that lie outside -1.96 and +1.96. This was 683 out of 13580, which represent less than 5% of the observations. To make sure that the there are no points that exert undue influence on the model, used the cooks distance to check the maximum of the cooks value which was 0.02637068 which is less than 1 and no cause of concern.

Since the models satisfy all the assumptions, looks like we have a better model, which is Model 2. This can be confirmed by using anova() function to compare the two models. Model 2 is a better model at F(13263) = 3757.

CONCLUSION

On analyzing the data around the Melbourne housing market, I have drawn a conclusion that location alone does not play a major role in describing the price for the house. Three major factors that play a role are location, number of bedrooms and the type of the house.
I can deduce that these three factors explain 76% of the variation in the price of the houses which is significant and should be considered while checking for houses. In addition, to put light on how we can predict the price for houses in a particular suburb and a constant type of house is - on increasing the number of bedrooms in a house by 1, the price increases by 0.81 – 0.85 times.