**Predictor table for y = adom**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predictor | Expected sign of effect on y | Rationale |
| 1. | List price | + | If a property is priced too high relative to its value or the local market, it may sit on the market longer and vice-versa. |
| 2. | Age | + | older properties may have a longer ADOM than newer properties due to factors such as outdated design features, wear and tear, or the need for significant renovations. |
| 3. | Sqft | + | Bigger house sizes could be more desirable than smaller house sizes. |
| 4. | Beds | + | Properties with more bedrooms may take longer time to sell as they can be expensive. |
| 5. | bathstotal | + | Like Number of bedrooms, having higher number of bathrooms is an indication of an expensive house that would take time to sell. |
| 6. | Pool | + | Having a private pool is a sign of an expensive house and could take longer time to sell. |
| 7. | Splsale | - | If it was a splsale, we can expect that it sells quickly because a short sale or a bank owned REO is a liability for the seller and must be sold quicker than other properties. |

|  |
| --- |
| * Slnoskm, status, address and subdivn uniquely identify each row and hence are not usable for predicting adom. * Bathsfull and bathshalf have not been considered because bathstotal alone could capture the total number of baths in the house. * Lotsqft has not been considered because cost per square feet \* sqft gives the price sold. Hence sqft is what the customers are paying for. * Roof and garages would not have any significant impact on adom. * List price per square foot has been captured by total list price and sales price per sq foot would not have any impact on adom. * The datesold is the date at which the property was sold, and it would not have any impact on adom.   **Feature engineering**-   1. To identify different drives among the neighborhood, we stripped off the numbers from the “address” column leaving just the drive name. Eg- “1634 Longrun river dr” which is a unique id to “Longrun river dr”. 2. Dropped 4 rows “NA” rows. 3. Converted garages, splsale to factors and relevelled according. 4. Converted roof to just 2 levels – “Shingle” and “Tile”. 5. Created an “age” variable: “Yrblt” – “Pending Date” 6. Converted pool levels to binary. “Private”: 1 and “Not Private”: 0   **Model 1**- Has all the variables from the predictor table. We notice that beds and sqft have a high correlation. Bathstotal and beds have a high correlation and listprice and sqft have a high correlation. Therefore, we drop beds, bathstotal and sqft from model 2.  **Model 2**- We notice that pool\_binary1 i.e. having a private pool has a very feeble impact on adom. This is the reason why it is not included in Model 3.  **Interpretation of the top 3 predictors of price from Model 3**-   * For each 100,000 $ increase in "listprice”, "adom" is expected to increase by 16 in days, keeping rest of the predictors constant. * For each one-unit increase in "age" (in years), "adom" is expected to increase by 4 days holding rest of the predictors constant. * Compared to properties with No sales type, properties owned by bank that are auctioned are expected to have a decrease of 53 days in "adom", holding rest of the predictors constant. * Compared to properties with no sale type, properties owned by Bank/REO are expected to have a decrease of 4 days in "adom", holding rest of the predictors constant. * Compared to properties with no sale type, properties that are sold as a “short sale” are expected to have an increase of 57 days in "adom", holding rest of the predictors constant. |
|  |

**Checking LINE assumptions on Model 3**-

Chart

Description automatically generated

We notice that none of the LINE assumptions are met. Let’s check for **Independence** using the Durbin-Watson Test.

Text

Description automatically generated

The Durbin-Watson statistic is 1.862033, which is closer to 2 than to 0 or 4. This value indicates that there is less evidence of autocorrelation. The p-value is 0.114, which is greater than 0.05. This means that there is not strong evidence against the null hypothesis that there is no autocorrelation in the data, and the evidence is not sufficient to conclude that there is autocorrelation in the data.

**Multicollinearity**- Does not exist.

Text

Description automatically generated

High VIF value for the listprice and the squared term which is expected. But the rest are less than 5 which is acceptable.

Chart, histogram

Description automatically generated

**Predictor table for y = sale price**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predictor | Expected sign of effect | Rationale |
| 1. | Lotsqft | + | Larger the area of a home, higher the price because of more living space. |
| 2. | Age | - | In general, newer properties tend to have a higher sale price than older properties because they are perceived to be in better condition and require less maintenance and repair work. |
| 3. | Garages | +/- | Homes with single garage might have lower price compared to two garages. But as garages get 2+, living space decreases and the SP would decrease. |
| 4. | Pool | + | Pools are a desirable amenity and would drive up the sales price. |
| 5. | Roof | + | Certain types of roofs are desirable and certain are not. |
| 6. | CDOM | - | As CDOM increases, the home stays on the market for a long time, indicating that it could be undesirable for various reasons and to sell it, the Selling price must be reduced. |
| 7. | Beds | + | More beds implies that the price could go up. |
| 8. | Bathstotal | + | Same reason as Beds. |
| 9. | Splsale | -/+ | If it was a special sale, the price would go down because it is either a short sale, and if it were in an auction, it could go up beyond its listing price. |

Rationale behind not including the other variables-

Slnoskm is considered to predict prices as they are more like row ids. Status of every house is “sold” and hence we won’t get any additional information from it to predict the sales price. Sqft is ignored because the buyer would pay for the whole lot and hence sqft would just contaminate the estimation. Cdom is considered in place of adom as cdom denotes the history of all days that the house has been unsold. “Yrblt” would be captured by the age variable. And in my opinion, having a Spa is not going to have any drastic effect on sales price.

Table

Description automatically generated

**Interpretation of the top 4 predictors of price from Model 3**-

* For each unit increase in "lotsqft", holding all other variables constant, "pricesold" is expected to increase by $16.
* For each unit increase in "beds", holding all other variables constant, "pricesold" is expected to increase by $36,000.
* For each unit increase in "garages", holding all other variables constant, "pricesold" is expected to increase by $36,000.
* roof2: This is a binary variable indicating whether the roof is tile or Shingle. The coefficient of $13,640 indicates that properties with tile roofs have an expected value of pricesold that is $13,640 higher than Shingle roofs, holding all other variables constant.

Chart

Description automatically generated

**Linearity**: There seems to be a slight upward curve in the graph and hence the linearity assumption is not met.

**Normality**: The curve seems to align with the red line for most of the part except the edges where it falls away. Normality assumption is not met. And the histogram also indicates there’s a slight skewness towards the right.

**Equality of variances**: This assumption is also not met because there seems to be a fanning out effect in the graph which we do not want to see.

Text

Description automatically generated with low confidence

**Testing for Independence**: The Durbin-Watson statistic is 1.454958, which is less than 2, indicating the presence of positive autocorrelation in the data.

**Multicollinearity**- Does not exist.

Text

Description automatically generatedChart

Description automatically generated

The VIF values are less than 5 indicating that there is not a problem of multicollinearity among the predictors.