# Predictive Analytics

#### **Part One—Focal Point and a Dependent Variable**

1. The decision or prediction to target using predictive analysis.

* I occasionally check home prices and it is interesting that despite the pandemic house prices have not declined. There are several factors that impact house value, such as location, vicinity of shopping centers, age, area income level, environment, politics etc. I want to find out if housing prices will continue to appreciate despite the pandemic economic crisis and will it be a smart decision to invest in a house in the near future. Perhaps an analysis here will help me get a clear picture from the USA housing dataset from Kaggle.

Target of analysis:

1. The dependent variable that will guide my prediction or decision.

* Dependent variable: The dependent variable in my analysis will be housing price. I want to find out how much independent variables such as income level, population, house age and # of bedrooms in a house affect house price.

1. Table: At least three independent variables that have association with the dependent variable. Each independent variable are identified as quantitative or categorical and discuss its expected impact on the dependent variable.

|  |  |  |
| --- | --- | --- |
| **Independent Variable** | | |
| **Summary of independent variable** | **Categorical or quantitative?** | **Argument for / description of the associates with the dependent variable** |
| Avg. Area Income | quantitative | Household income has been one of the determinants of prices in the market. If it is positively correlated with house price, it will be interesting to find out how strong the relationship is. |
| Avg. Area number of Bedrooms | quantitative | Number of bedrooms is a factor in house price, but will it be more important than income? |
| Avg. Area population | quantitative | It is unclear how population would be related to price; the analysis will shed light in this regard. |
| Avg. Area House Age | quantitative | Newer houses might be more expensive, but old is gold. The analysis will show if newer houses are more expensive or not. |

**Part Two—Map Decisions to Outcomes**

Map how the decision or predicted outcome is related to the independent variables.

|  |  |  |
| --- | --- | --- |
| **Candidate Independent variables** | | |
| **Independent variable** | **Regression equation** | **Screenshot of scatterplot** |
| Avg. Area Income | y = 2.0376x + 4.2792 |  |
| Avg. Area House Age | y = 1.5649x + 4.0917 |  |
| Avg. Area Population | y = 1.1322x + 2.2336 |  |
| Avg. Area Number of Rooms | y = 0.7173x + 6.137 |  |

* Average Area Income has the strongest positive relationship with price compared to other independent variables. It is generally the case that house prices are higher in areas where households have higher income. The intercept and slope are also positive which means that as income increases, price also tends to increase. So, the scatter plot makes sense and the regression equation and R2 supports this claim.
* The scatter plot and regression equation show that House Age has a positive relationship with price. In addition, the relationship between house age and price is stronger than the latter two independent variables. I believed that newer houses might be more expensive than older ones. The analysis in part 1 shows that as houses get older, their prices go higher which is counterintuitive.
* Average Area Population also has a positive relationship with price. The regression equation and R2 are also positive. Although this variable has a weaker relationship than the above two, but it makes sense. There tends to be more demand for houses in populated areas and demand drives price in Capitalist markets.
* Average Area Room Number has a positive relationship with price, but it is the weakest amongst the 4 independent variables. The slope and intercept are positive which shows that price seem to increase based on the number of rooms. I believed that larger houses might be more expensive and therefore will have a stronger positive relationship after Avg. Area Income, but the analysis show a weaker relationship.

1. Graphical user interface, application, table, Excel

   Description automatically generatedA multiple regression for my data.

**Part ThreeGenerate a Revised Regression Equation**

Here I have further refined my regression model potentially by transforming variables for nonlinear relationships and/or including or excluding variables from the regression to address multicollinearity.

Addressing nonlinear relationships:

Graphical user interface, application, table, Excel

Description automatically generated

* All relationships in part 1 and above residual plots show linear relationships.

1. The Semilog and Log-log Transforms tool are used to create and transform scatterplot for each independent variable listed.

* Although all relationships in part 1 are linear, I still wanted to use log transform to find out if I could further refine the plots. R2 does change in all but not a lot.

1. The independent variables with nonlinear relationships.

|  |  |  |
| --- | --- | --- |
| **Possible nonlinearities** | | |
| **Independent variable** | **Transform used (log or semilog)** | **Screenshot of transform plot** |
| Avg. Area Income | log |  |
| Avg. House Age | Log |  |
| Area Population | Log |  |
| Avg. Area Number of Rooms | Log |  |

Addressing Multicollinearity

1. A correlation table for my independent variables.

Graphical user interface, application, table, Excel

Description automatically generated

* There are no independent variables which could be a source of multicollinearity. The Correlation matrix shows that each independent variable has a stronger relationship with Y (Price).

**Part Four—Validate Model**

According to forbes.com, one clear beneficiary of the housing supply and demand imbalance in the market amidst the pandemic crisis will be price appreciation. Price is expected to remain strong. Home buyers will likely feel pressure to escalate their bids to win the “bidding war” on homes that are for sale. As such, the variable of interest in this project will be house prices. The four independent variables like; average area income, average house age, average number of rooms and average area population are amongst the key determinants of price in the housing market. Other important independent variables, which are not covered in this analysis, can be location, vicinity of shopping centers, access to main roads, or price correction by the government.

Per the above scatter plots, regression equation, R2 values, regression analysis and correlation matrix, there is a strong positive relationship (y = 2.0376x + 4.2792, R2 = 0.4856) between average area income and price, indicating that house prices are higher in areas where people with higher income live. Average area population also has a positive relationship (y = 1.1322x + 2.2336) which is an indicator of higher demand for houses in populous areas. A positive relationship (y = 0.7173x + 6.137) of average number of rooms and price show that generally larger houses are more expensive. I intuitively believed that house age might have a negative correlation with price, but the analysis here proved me wrong showing a positive relationship (y = 1.5649x + 4.0917).

The scatter plots, regression equation, R2 values and correlation matrix model a good analysis without any non-linearities or multicollinearities. The strong R2 and the p-values which are smaller than 0.05 further support the good analysis. Furthermore, the coefficients of independent variables in the regression analysis are all positive, but the intercept.

The negative intercept is nonsensical, indicating that there is a negative relationship between independent variables and the dependent variable (price). I tried to revisit the model by removing one of the independent variables with a weak relationship, as well as by dividing price with number of rooms to create a new independent variable. Despite several efforts, every regression model showed a negative intercept. Perhaps, a larger sample, more than 250 rows of data, is needed to fix this error. Or even though there a positive relationship between the independent variables and the dependent variable, there is an overall weak linear relationship with a negative intercept which does not strongly support the model. In conclusion, the negative intercept implies that the independent variables in this model are not the main drivers of house price, and that further research and analysis is required to find out the principal determinants of house price to justify smart decision making.