CSC 423

Appraising Residential Properties To Boost Sales and Win Over Clients - Technical Report

Team: PredictThis!

Abstract

A dataset was obtained from Kaggle with history of commercial real estate properties being sold from the 1990's to 2018. The data was assessed by creating several different models which were carried out by each team member individually. Finally, the models were compared to be able to obtain worthy information to our findings in terms of predicting the price of real estate in Washington, DC.

Introduction

In today's world real estate has a very large impact on the economy and vice versa. One of the crucial aspects of the real estate business is proper appraisal of property value. For large real estate companies or even for the government to be able to place taxes effectively they would need to have a very good understanding to how these properties should be valued and also how their value might change in the future. These properties are valued based on several aspects of each property some of which do not have a significant effect on the properties value. D.C. Geographic Information System has obtained a large dataset on properties sold in Washington D.C. dating back to the early 1990's. Our team will try to find an effective regression model to predict the price of this housing effectively and find out which model is best.

METHODOLOGY

For this project, a complex and up to date dataset is selected from: https://www.kaggle.com/christophercorrea/dc-residential-properties. This dataset is provided by D.C. Geographic Information System for many housings in Washington, D.C and includes useful information on houses sold in Washington, D.C. Using this dataset, PRICE of the house in Washington D.C. area is predicted using Linear Regression Technique. The original raw dataset contained 49 columns and 159,000 observations, however 2000 observations are randomly selected using Excel formula. Duplicate and missing values are removed in Excel. Out of 48 independent variables, initially 19 are chosen for my individual analysis based on housing research and domain knowledge. Then I imported the data into SAS and performed data cleaning and pre-processing using SAS code. Creating new variables, recoding, binning, dummy variables are all done using SAS code. My methodology includes following processes: 1) Pre-processing steps 2) Exploration steps 3) Analysis steps 4) Fit the full model check for all diagnostics 5) Split the data into Train/test 6) Model selection using training set 7) Test the performance using test set 8) Compare performance and 9) Compute predictions.

ANALYSIS, RESULTS AND FINDINGS

Dependent variable (DV): 1 common dependent numerical variable

- *Price*: the price of the most recent sale

Independent variables (IV): Out of 48 independent variables, initially 19 are selected for my individual analysis and those variables are:

- 1 date/time variable: *SALEDATE*
- 1 location variable: *QUADRANT*
- 10 quantitative or numerical variables: *BATHRM, HF_BATHRM, NUM_UNITS, ROOMS, BEDRM, STORIES, GBA, KITCHENS, FIREPLACES, LANDAREA*

7 qualitative or text variables: *AC, QUALIFIED, STYLE, STRUCT, CNDTN, EXTWALL, QUADRANT*

Data Cleaning and Pre-Processing:

- After carefully looking at the dataset, 2000 observations are randomly selected for project out of 159,000 observations. Duplicate and missing values are removed and checked for data inconsistencies.

- Then dataset named "UMAIR_DC_Properties_Revised.csv" is imported into SAS using import statement. The initial dataset is named in SAS as out=house and it contains 49 columns with 2000 observations.

- After importing the dataset in SAS, 29 variables which have not been chosen for my individual analysis are dropped. Dataset with reduced # of columns is stored in new SAS dataset house_cleaned.

New Variables / Binning Variables:

Using SAS Code, following variables are converted and reduced into a smaller number of categories based on a different logic for each x-variable. All logics are based on the housing dataset research and domain knowledge. *See Appendix B.1 for output*.

SALEDATE:

o Only YEAR part is extracted from the date/time variable (SALEDATE) and stored into a new variable named (year). Then a new bin variable is created named (bin_SALEDATE) where:

- *bin_SALEDATE*=1 or Before Burst if year of sale is between 1992 to 2005 (These are the sale of homes before Bubble Burst in housing market of USA.

- *bin_SALEDATE*=2 or After Burst if year of sale is between 2006 to 2012 (These are the sale of homes after Bubble Burst in housing market of USA.

- *bin_SALEDATE*=3 or Recent Sales if year of sale is between 2013 to 2018 (These are the most recent years sales of homes.

STYLE:

o A new bin variable is created as (bin_STYLE) where:

- *bin_STYLE*=1 or Single Story if style of the structure in ('1 Story', '1.5 Story Fin', '1.5 Story Unf').

- *bin_STYLE*=2 or Double Story if style of the structure in ('2 Story', '2.5 Story Fin', '2.5 Story Unf').

- *bin_STYLE*=3 or Triple Story if style of the structure in ('3 Story', '3.5 Story Fin').

- *bin_STYLE*=4 or Other Style if style of the structure in (‘4 Story’, ‘Bi-Level’, ‘Default’, ‘Split Foyer’, ‘Split Level’).

STRUCT:

o A new bin variable is created as (bin_STRUCT) where:

- *bin_STRUCT*=1 or Single Family if type of the structure is (‘Single’’).

- *bin_STRUCT*=2 or Multi Family if type of the structure is (‘Multi’).

- *bin_STRUCT*=3 or Town Home if type of the structure in (‘Row End’, ‘Row Inside’, ‘Semi-Detached’, ‘Town End’).

CNDTN:

o A new bin variable is created as (bin_CNDTN) where:

- *bin_CNDTN*=1 or Not Good Condition if the condition of the structure in (‘Average’, ‘Fair’).

- *bin_CNDTN*=2 or Good Condition if the condition of the structure in (‘Excellent’, ‘Very Good’, ‘Good’).

EXTWALL:

o A new bin variable is created as (bin_EXTWALL) where:

- *bin*_EXTWALL=1 or Brick if the type of exterior wall in (‘Brick Veneer’, ‘Brick/Siding’, ‘Brick/Stone’, ‘Brick/Stucco’, ‘Common Brick’, ‘Face Brick’).

- *bin*_EXTWALL=2 or Stone if the type of exterior wall in (‘Stone’, ‘Stone Veneer’, ‘Stone/Siding’, ‘Stone/Stucco’, ‘Stucco’, ‘Stucco Block’).

- *bin*_EXTWALL=3 or Frame if the type of exterior wall in (‘Hardboard’, ‘Shingle’, ‘Vinyl Siding’, ‘Wood Siding’).

- *bin*_EXTWALL=4 or Other if the type of exterior wall in (‘Aluminum’, ‘Concrete’, ‘Concrete Blo’, ‘Metal Siding’).

Dummy Variables:

All the qualitative data is examined and following dummy variables are created for the data to be able to fit a regression model later: *See Appendix B.2 for output*

AC	dumAC
Y	1
N	0

QUALIFIED	dumQUALIFIED
Q	1
U	0

QUADRANT	dumNW	dumSE	dumSW
NE	0	0	0
NW	1	0	0
SE	0	1	0
SW	0	0	1

Bin_SALEDATE	dumAfterBurst	dumRecentSales
1	0	0
2	1	0
3	0	1

Bin_STYLE	dumDoubleStory	dumTripleStory	dumOtherStyle
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

Bin_STRUCT	dumMultiFam	dumTownhome
1	0	0
2	1	0
3	0	1

Bin_CNDTN	dumGoodCndtn
1	0
2	1

Bin_EXTWALL	dumStone	dumFrame	dumOtherExtwall
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

Means, Descriptive, Frequencies:

Under this section, descriptive and frequencies are checked for y and all x-variables and extreme observations are removed from the data.

The total number of observations shows 2000 for all variables and there no missing values.

For dependent variable, PRICE, the lowest price of the most recent sale voted is $23,000. 25% of the data falls below the first quartile which is $318,500. The midway point of the data is $556,500, 50% of our data falls below $556,500. 75% of the data falls below $822,500 and the maximum price of the most recent sale of house is $8,500,000. The Interquartile Range (IQR) is $504,000. The middle fifty will be very useful for us when we are doing the prediction or regression analysis. *See Appendix B.3 for output*.

After running descriptive for all variables, 75 observations are found as outliers and influential points. *See Appendix B.4 for output.*

Those 75 observations are removed from the dataset and new data with 1925 observations are saved into new dataset "houseNew_1".

After removing outliers and influential points, frequencies are checked for all variables on new dataset houseNew_1. *See Appendix B.5 for output.* The cumulative frequencies for all the variables show total 1925 observations. Frequency and Percent columns are indicating the distribution of observations among different categories and it seems that in most of the variables, categories are not equally distributed.

Histograms with Normal Curves and 5 number summaries:

See Appendix B.6 for output

The distribution for dependent variable PRICE has a longer right tail and it does not seem normal. The histogram shows skewedness towards right and there is big difference between mean ($651,695) and median ($560,000). Histograms are indicating that there are couple of outliers and influential points. The skewness in the other x-variables such as (BATHRM, HF_BATHRM, NUM_UNITS, ROOMS, BEDRM, STORIES, KITCHENS, and FIREPLACES) are due to natural variation in the dataset as these variables have values such as 1,2,3,4,5…. Etc. But we cannot use them as text or dummy variables because these are the number of counts and must be treated as numerical variables.

Transformation for PRICE:

See Appendix B.7 for output

Due to the very high right tail, there is no significant difference noted in the skewedness of the PRICE variable even after removing outliers and influential points. Therefore, those observations are not removed from the dataset and a log transformation is applied on PRICE variable. A new variable is crated as "ln_PRICE" and this is going to be used from now onwards in analysis of the report.

Histogram for Transformed variable ln_PRICE:

See Appendix B.7 for output

After log transformation of PRICE variable, the distribution of ln_PRICE looks almost symmetric and normal. There is very small difference between the mean ln_PRICE (13.12) and the median ln_PRICE (13.23). The minimum ln_PRICE value is 10.46 and the maximum ln_PRICE value is 15.60.

Scatterplots and Correlation:

By looking at the GPLOTS (*See Appendix B.8 for output)* and Scatterplots Matrix (*See Appendix B.9 for output)*, there seems to be a positive linear association between all pairs except NUM_UNITS. The following correlation values (*See Appendix B.10 for output)* also support these conclusions that there is a positive linear association among these pairs of variables except NUM_UNITS. We can see some dots on the far-right top of GBA and LANDAREA which we need to address in our analysis. They can possibly be the outliers or influential points. For variable GBA, most of the dots are clustered around the lower band on x-axis and middle band on y-axis. For variable LANDAREA, majority of the dots are clustered around the very lower band on x-axis.

The GPLOTS for other numerical x-variables such as (BATHRM, HF_BATHRM, NUM_UNITS, ROOMS, BEDRM, STORIES, KITCHENS, and FIREPLACES) are showing straight lines on the graphs, this is because of the natural variation in the dataset as these variables have values such as 1,2,3,4,5…. Etc. These are # of counts of Rooms, kitchens, fireplaces etc.

However, we cannot check correlation, linearity, and normality using correlation values and scatterplots for dummy variables because it has values nothing but 0 and 1. With dummy variables, we don't see a linear relationship, points are scattered around 0 and 1. We see nothing but vertical lines on 0 and 1. We can just use dummy variables in regression analysis and see what happens. We can use standardized estimates to see which predictors are important.

x-variables	In_PRICE	Degree of Correlation
GBA	0.46857	Moderate positive association
LANDAREA	0.19959	Weak positive association
BATHRM	0.45856	Moderate positive association
HF_BATHRM	0.28565	Weak positive association
NUM_UNITS	0.02535	Very weak positive association (almost 0)
ROOMS	0.27988	Weak positive association
BEDRM	0.30228	Weak positive association
STORIES	0.26097	Weak positive association
KITCHENS	0.08204	Very weak positive association
FIREPLACES	0.45928	Moderate positive association

Multicollinearity:

Correlation values larger than 0.9 or so among independent variables indicate a serious collinearity problem. After checking the output of Pearson Coefficient Correlation (*See Appendix B.10 for output)*. NUM_UNITS and KITCHENS are correlated with each other. The correlation value of 0.919 indicates a Perfect multicollinearity. X-variable "NUM_UNITS" is dropped because of the multicollinearity with KITCHENS and a very low (almost 0) correlation with dependent variable In_PRICE.

Full Model and Residual Plots:

Overall Test on Goodness-of-Fit:

(*See Appendix B.11 for output)*

Null hypothesis: None of the x-variables included in the model have any association with Y-variable (In_PRICE).

Alternative hypothesis: H_a: At least one coefficient $\beta_j \neq 0$

At least one x-variable has a significant effect on changes in Y-variable (In_PRICE).

F-value is 146.28 and the p-value associated with F-value is much smaller then alpha=0.05. Therefore, the null hypothesis of no association between y-variable (In_PRICE) and x-variables is rejected. The F-test gives strong support to the fitted model and In_PRICE can be explained by model.

RMSE 0.4425 is lower.

Adj-R^2 = 0.6537 means 65% of the variation in In_PRICE can be explained by the model. The remaining 35% of the variation in In_PRICE remains unexplained. 0.65 ADJ-R2 is pretty good for the model.

Significance:

Following x-variables have p-value > 0.05: *(LANDAREA, ROOMS, BEDRM, STORIES, dumSW, dumAfterBurst, dumRecentSales, dumMultiFam, dumGoodCndtn).*

These variables are insignificant, and we should remove them from the model one by one and re-check the p-values. However, this will be done using different Model Selection Methods during the training process on split dataset.

Model Assumptions:

Linearity: Scatterplots and normal quantile plots indicating linearity.

Constant Variance and independence:

Studentized Residuals plot against the predicted values (*See Appendix B.12)* show some problems of constant variance with points not randomly scattered around the zero line. Pattern of the residuals decrease as the predicted values increase. It shows decreasing funnel shape pattern, violating independence assumption. Generally Constant variance and independence are related. If one is violated, then other one is too. Residual plots for independent variables GBA and LANDAREA (*See Appendix B.14)* also indicating the constant variance issue with decreasing funnel shape pattern.

Normality of errors: Normal probability plots (QQ Plot) *(See Appendix B.13)* show normality as the points lie close to the line. It has almost 45-degree line. However, if we check the residual vs quantile plot under diagnostics (*See Appendix B.13)* so there are couple of dots in the tails. Those are outliers and influential points which needs to be addressed.

Outliers and Influential Points:

Studentized Residuals plot against the predicted values (*See Appendix B.12)* shows a lot of observations with studentized residuals > +-3 which are outliers and influential points.

Studentized Residuals and Cook's D for In_PRICE (*See Appendix B.15)* also indicating a lot of observations with red (outlier) and blue (Influential Point) arrowheads.

Actions to take about Outliers and Influential Points:

- Examine the adj-R2 value = 0.6537.
- Remove the observations with red (outlier) and blue (Influential Point) arrowheads first.
- Re-Check the adj-R2 value, residual plots, and p-values of the predictors to see if they improved. If it doesn't, keep it as part of your observations.

Re-checking assumptions after removing outliers/influ points:

After removing 44 observations (outliers and influential points), following model improvement is noted:

- Adj-R2 improved from 0.6537 to 0.6982. (*See Appendix B.17)*

- F-value also improved from 146.28 to 174.96 and is very high. (*See Appendix B.17)*

- F-statistic is less than 0.05. Overall goodness of fit test shows that at least 1 predictor is significantly associated with Y. (*See Appendix B.17)*

- The residual plots show constant variance, independence, normality and linearity. (*See Appendix B.18 and B.19)*

Model Validation / Training and Testing:

I am splitting the cleaned dataset (houseNew_5) using a random seed into a training and test set.

Training Set: Created using at least 75% split to estimate/fit the model. (*See Appendix B.20).*

Testing Set: Created using an at least 25% split to test the predictive performance of the model.

Goal: The main goal is to test how well the model predicts new data (out of sample).

Step-1: Estimate model using training set:

Two model selection methods (STEPWISE and ADJRSQ) are used on training dataset to estimate the models.

STEPWISE model selected method selected 13 predictors based on alpha value < 0.01. Following are the predictors selected: (*See Appendix B.21)*

(GBA LANDAREA BATHRM HF_BATHRM FIREPLACES dumAC dumQUALIFIED dumNW dumSE dumAfterBurst dumRecentSales dumMultiFam dumGoodCndtn.

ADJRSQ model selected method selected 18 predictors based on adjusted r2 values. Following are the predictres selected: (*See Appendix B.21)*

(GBA LANDAREA BATHRM HF_BATHRM BEDRM KITCHENS FIREPLACES dumAC dumQUALIFIED dumNW dumSE dumSW dumAfterBurst dumRecentSales dumMultiFam dumGoodCndtn dumStone dumFrame).

Step-2: I computed predictions on testing set and compared them with observed values. An adequate model is producing predictions that are close to the observed values. (*See Appendix B.21)*

Overall, which model is better?

Model-1: 13 predictors	**Model-2: 18 predictors**
Train RMSE: 0.40230 R^2: 0.6894 Adj-R^2: 0.6865 GOF: OK Residuals: OK	**Train** RMSE: 0.40014 R^2: 0.6938 Adj-R^2: 0.6899 GOF: OK Residuals: OK
Test RMSE: 0.40748 MAE: 0.32102 R^2: $yhat^2 = 0.84554^2 = 0.71493$ Adj-R^2: $\lvert model\ R^2 - R^2_{cv} \rvert = 0.02553$ CV-R^2: OK < 0.3	**Test** RMSE: 0.40527 MAE: 0.32024 R^2: $yhat^2 = 0.84726^2 = 0.71784$ Adj-R^2: $\lvert model\ R^2 - R^2_{cv} \rvert = 0.02404$ CV-R^2: OK < 0.3

Overall, Model-1 is better as it has a smaller number of predictors (13) as compare to Model-2 with (18) predictors. RMSE, MAE, R2, Adj-R2, and CV-R2 are almost same for both models. The residual plots show constant variance, independence, normality and linearity for both models. (*See Appendix B.21).*

Final Fitted Model:

(*See Appendix B.22)*

GENERAL EQUATION

$$\mathbf{In\_PRICE = B_1 + B_2*dumRecentSales + B_3*dumAfterBurst + B_4*GBA + B_5*FIREPLACES + B_6*dumNW + B_7*dumQUALIFIED + e}$$

Where dumRecentSales=1 if bin_SALEDATE='3', and dumRecentSales=0 if bin_SALEDATE=1

Where dumAfterBurst=1 if bin_SALEDATE='2', and dumAfterBurst=0 if bin_SALEDATE=1

Where dumNW=1 if QUADRANT='NW', and dumNW=0 if QUADRANT=NE

Where dumQUALIFIED=1 if QUALIFIED='Q', and dumQUALIFIED=0 if QUALIFIED='U'

FITTED REGRESSION EQUATION

$$\mathbf{In\_PRICE = 11.28 + 0.87*dumRecentSales + 0.60*dumAfterBurst + 0.0003*GBA + 0.22*FIREPLACES + 0.34*dumNW + 0.42*dumQUALIFIED + e}$$

Where dumRecentSales=1 if bin_SALEDATE='3', and dumRecentSales=0 if bin_SALEDATE=1

Where dumAfterBurst=1 if bin_SALEDATE='2', and dumAfterBurst=0 if bin_SALEDATE=1

Where dumNW=1 if QUADRANT='NW', and dumNW=0 if QUADRANT=NE

Where dumQUALIFIED=1 if QUALIFIED='Q', and dumQUALIFIED=0 if QUALIFIED='U'

dumRecentSales: Assuming all other variables constant, PRICE of the house increases by $138,690 for the houses sold during 2013 to 2018 (Recent Sales) as compare to houses sold during 1992 to 2005 (Before Bubble Burst). Computed as $100*(e^{0..87} -1) = 138.69$ x 1000 = $138,690.

dumAfterBurst: Assuming all other variables constant, PRICE of the house increases by $82,210 for the houses sold during 2006 to 2012 (After Bubble Burst) as compare to houses sold during 1992 to 2005 (Before Bubble Burst). Computed as $100*(e^{0..82} -1) = 82.21 \times 1000 = \$82,210$.

GBA: Assuming all other variables constant, for any 1 sqft increase in GBA (Gross Building Area), PRICE of the house increases by $30. Computed as $100*(e^{0..0003} -1) = 0.030 \times 1000 = \$30$.

FIREPLACES: Assuming all other variables constant, for any 1 number increase in FIREPLACES, PRICE of the house increases by $24,600. Computed as $100*(e^{0..22} -1) = 0.24.60 \times 1000 = \$24,600$.

dumNW: Assuming all other variables constant, PRICE of the house increases by $40,490 for the houses located within Quadrant of NW as compare to houses located within Quadrant of NE Computed as $100*(e^{0..34} -1) = 40.49 \times 1000 = \$40,490$.

dumQUALIFIED: Assuming all other variables constant, PRICE of the house increases by $52,190 if the customer qualified for the loan as compare to not qualifying for the loan. Computed as $100*(e^{0..42} -1) = 52.19 \times 1000 = \$52,190$.

A model with a smaller number of predictors is better. Based on the 13 predictors selected by STEPWISE model selection method above, I picked 6 important predictors which has the highest influence on Price of the homes.

By looking at the Standardized parameter estimates, following predictors have the largest absolute value of the standardized coefficient representing that it has the greatest influence on PRICE and can be considered the strongest predictor of housing price: dumRecentSales, dumAfterBurst, GBA, FIREPLACES, dumNW, dumQUALIFIED

Other predictors have very small standardized coefficient values. Not having those predictors in the model will have very minimum effect as these are not the strongest predictors of PRICE.

F-value is 567.94 and the p-value associated with F-value is much smaller then alpha=0.05.

RMSE 0.4353 is very low.

Adj-R^2 = 0.6440 means 64.5% of the variation in In_PRICE can be explained by the model. The remaining 35.5% of the variation in In_PRICE remains unexplained. 0.6440 ADJ-R2 is pretty good for the model.

Significance: All predictors have p-value < 0.0001

Constant Variance, independence, and Normality: (*See Appendix B.23).*

The residual plots of final fitted model show constant variance, independence, normality and linearity. There are no issues found in plots.

FUTURE WORK

The most important factor which has been discovered so far in this research is the house prices effects due to recession and housing bubble burst. (According to Nihar Bhagat "In today's real estate world, it has become tough to store such huge data and extract them for one's own requirement. Also, the extracted data should be useful. In-depth details of every property will be added to provide ample details of a desired estate. This will help the system to run on a larger level. "[1]). In future, the model dataset can be analyzed by using non-linear regression techniques. Also, transformations of some of the numerical x-variables can also help better the results. Due to time constraints and limited knowledge, not many location-based x-variables were used in this research although location of the property is one the biggest factors in price predictability.

REFERENCES

[1].": Bhagat, Nihar. 2016. House Price Forecasting using Data Mining. *International Journal of Computer Applications (0975 – 8887)*