Aaron Niecestro
Kingsley Iyawe
Generalized Linear Models
Project Report

DC Properties Qualification Report

Introduction

This report is about using binary logistic regression to figure out which variables and their effect those variables have on what makes a property qualified to sell. Qualified property means that the paperwork needed to sell a house, the deed of the property, approval from banks (if needed), etc. is completed, and the property inspection is passed. The reason this type of study is being conducted is that my partner and I both go to American University which is in the District of Columbia. We thought and believed that since a lot of students live in either DC or one of the surrounding states (West Virginia, Virginia, Maryland), this would be something we could analyze and learn from. Also, we felt that maybe some of our fellow students will be property owners or apartment renters in the coming future, if not already, and this would give insight into whether they will be able to pick a qualified and right place for themselves to live.

The next step was finding data that we could use for a logistic regression model. We found our data relatively fast from Kaggle D.C. Residential Property, and agreed upon using binary logistic regression analysis, although we could have also used nominal multinomial and ordinal logistic regression by using a different response variable then qualification. Once the data and type of logistic regression analysis were decided, the next step was coming up with questions we wished to answer. The questions we created and tried to answer were as follows: 1) What does the Qualification column in the dataset mean? 2) What qualifies a residential property to be sold on the housing market? 3) Is the property pricing the most important factor in determining whether a property is qualified to go on the market? 4) Do the realtors even care about whether a

property is qualified to sell before listing it or is it all about the money? 5) Are we creating the most optimal regression for modeling properties? 6) Do we follow previous linear regression housing model approaches for predictor variables, or should we come up with our own model and approaches from scratch? and 7) Is money the most important thing? If so how does that define the world? With these questions in mind, we started to clean, assess, and manipulate the data, so no extremes were used in the analysis and modeling processes.

Data

Following the download of the data we started to assess and figure out what each column represented and the importance of each column. The original data had 158957 rows and 49 columns. To do this we had to read the description of the columns on the Kaggle site and google what we might not have known since this is not our area of expertise. In the beginning, we ran into a slight problem with not knowing what the qualification column in the original dataset meant. To resolve this problem of ours, we tried to reach out to the original uploaders of this dataset, but the original uploaders have not gotten back to us yet. So, we researched ourselves what a qualified property might be and the things a person should do to sell their property. The research later becomes what the qualification response variable description.

Since the dataset had 49 columns, this report will describe only the variables used in the models and analysis. This is because it will take up too many pages otherwise. The model variables were as follows: 1) PRICE, price of most recent sale, 2) BATHRM, the number of full bathrooms, 3) HF_BATHRM, the number of half bathrooms (no bathtub or shower), 4) AC, whether the property has air conditioning, 5) ROOMS, the number of Rooms, 6) BEDRM, the number of bedrooms, 7) STORIES, the number of stories in the building or property, 8) QUALIFIED, whether a property is qualified to sell, 9) STYLE, the style of the property, 10)

CNDTN, a verbal rating of the condition of the property, 11) KITCHENS, the number of kitchens, 12) FIREPLACES, the number of fireplaces, and 13) WARD, the ward and the ward number (District is divided into eight wards, each with approximately 75,000 residents). Although there were more variables, these variables were not used in our model and shall be described in a later report.

Unfortunately, we started to have a lot more issues even before the cleaning process began. One of the key issues we noticed right away and were coming across was that a lot of data was missing in most the columns. Each column besides the ID column had missing rows between 1 observation to over fifty percent. We also had two columns (complex number and living GBA) that had to be taken out since they had no data entries in any of the rows. We come to the decision not to add the data which could have been found on realtor sites that had similar qualities. We did not fill in the blanks for the missing data because it would increase bias dramatically and who knows whether the data we could have added it would be the correct data. The executive decision we came down to was taking out all the blanks from our dataset and working with only the data that was downloaded.

Although this method was working great, we ran into some more problems. Some of these issues we were facing were data being entered either incorrectly, data having errors, and values incorrectly labeled. One example air conditioning column (AC). The AC column was supposed to be Y, yes, and N, no, but it had a third value of 0 which had to be later changed to N. Once the data cleaning was completed, we decided the bounds we wished to use for our analysis. The bounds we came with were rather long but necessary in lowering bias. The bounds we came up with were as follows: Price between $10,000 and $1,000,000, Fireplaces less then 8, kitchens less than or equal to 10, rooms less than 26, bedrooms less than 20, stories of your building less

than 100, bathrooms greater than 0, and half bathrooms great than 0. With these bounds in mind, we started to use visualizations to see what kinds of graphs we could create and data we are working with. Although this worked great for our visualizations, the dataset we used for a model had only the essential variables we wish to use. So, in total there were three datasets for this project which were called in order, the original dataset, the visualization dataset, and the model dataset. The original dataset as stated above had 158957 rows and 49 columns. The visualization dataset which used the specified bounds had 17522 rows and 46 columns. The model dataset which used the specified bounds had 33671 rows and 32 columns. Now that the data cleaning was completed we can move onto the analysis section where we will report on how we compiled our model and the model process. It should be noted that we were not happy and tried to figure out ways to get more than 1/5 of the original dataset to no avail.

Analysis

To complete our analysis, we tried to use our analysis processing skills. Before we conducted any analysis, we decided to break the model dataset into two parts called the training dataset and the validation dataset. The training dataset included approximately sixty percent of the models' datasets. The validation dataset included approximately forty percent of the models' dataset. The analysis processes in order were composed of model building, diagnostic processes, stepwise AIC, stepwise BIC, hypothesis testing which included Likelihood ratio test and goodness of fits tests, rebuilding the model, checking for multicollinearity issues, fixing multicollinearity issues, creating interaction terms, and creating the ROC Curve, and more diagnostic processes.

The first step in our analysis process was creating a lot of diagnostic plots and noting our observations of these plots. From these plots, we could come to a few conclusions. These

conclusions were as follows: 1) our data is not normally distributed so we will have to use transformations, 2) we have multicollinearity, so we will have to create variance inflation factor graphs and numbers to fix this issue (refer to table 1) we will need to apply transformations to our model. With these things in mind, we moved on to the model building process.

The second step in our analysis process was to build some models and choose which predictor variables we wish to use for our model. The first model we built was very simple but it had our basic requirements on what we believed was necessary at that time to model qualification. Our first model was using the qualification as the response variable and price as our only predictor variable. In the beginning, we believed that price is the most important variable and it should not be eliminated from our model because most housing model price as the response so we should have at a minimum price as a predictor variable. The basic model equation was Log (π/1-π) = 0.6811 + 0.000001726*Price and the AIC was equal to 17,988. So, for every additional pricing dollar, the odds a property being qualified to sell increases by a factor of 1.000002. When we tried to use the likelihood ratio test with a null hypothesis that β1 is equal to 0. We did this to determine if the price was supposed to be in the model going forward or not. Although if the result was not to reject the null hypothesis, we might have just noted it and continued with the analysis keeping price as a predictor variable anyway. Thankfully though the results stated to reject the null hypothesis and keep the price as a predictor variable.

Moving on we created a new model with all the predictor variables we felt were necessary. This new model included the price, the number of bathrooms, the number of half bathrooms, having air conditioning dummy variable, number of stories the property has or the building the property is in, the type of style of the property (14 categorical variables), the condition of the property (4 categorical variables), the number of kitchens, the number of

fireplaces, and the ward number where the property was located (ward 1 – 5). It should be noted that although there were originally 8 wards, after the data cleaning and creating the training and validation datasets we ended up with only 5 wards. This binary logistic regression model had an AIC equal to 17,410. From this model, we could tell we were on the right track in our model building process since the AIC decreased by 578 from our first basic model, but the further analysis still needed to be completed since this AIC was still very high.

The next step was figuring out if we could reduce the 32-betas in our model. To reduce the model, we used stepwise AIC and BIC. The stepwise AIC and BIC results showed us that we should keep the following predictor variables: price, bathrooms, AC=Yes, bedrooms, 14 Style dummy variables, 5 condition dummy variables, kitchens, and the 4 ward dummy variables. My partner and I made sure to double check these results with the likelihood ratio test by finding G-squared and p-value to make sure these variables that were being taken out were correct, and there were no other variables we missed taking out before moving on. The reason that we were so meticulous with getting rid of variables is that we wished to eliminate any cases of hidden multicollinearity, and our belief that having too many variables would disrupt the model. It should be noted that before moving on we decided to take out the style dummy variables from our model, even though it slightly increased the AIC. This is because the style was creating too many betas, majority of style dummy variables were statistically insignificant, and they were also making the rest of our predictor variables be statistically insignificant.

Once the final single predictor variables were selected we decided to create all possible 2-way interactions terms. We could have created interaction terms on our own from what we felt was the most important, but we did not wish to miss any type of interaction terms that could have been beneficial to creating a better binary logistic regression model for qualification. From this

2-way interaction model, we used stepwise AIC, BIC, and likelihood ratio tests to determine which variable and interactions terms because these variables would become the founding base for our final model with transformations. The variable and the interaction terms that were created from all this analysis were as follows: price, AC=Yes, bedrooms, 5 condition dummy variables, the 4 ward categorical variables, price times AC=Yes, price times rooms, and price times the 4 ward categorical variables.

Following the creation of our interaction terms model, we looked back at the diagnostic plots and created some more diagnostic plots to see if we if the underlying concerns we had were still around. It seemed that our data was still not normally distributed so we would have to create some transformations to our interaction terms model. The good thing was that multicollinearity we noticed and were concerned about was no valid. However, we did notice that there were some outliers in training set data, so we checked to see whether they were significantly impacting our model and if they were significantly impacting our model we took them out. Although because the training set dataset were we working was large, we might have missed taking out outliers.

The next step was to add some transformations to our model to make the data and our model more normally distributed. We found that although we could add transformations to the single predictor variables, when we applied these transformations to the interaction terms, the AIC and BIC numbers were increased as a result and some variables were becoming statistically insignificant. So, with the little option left, my partner and I left the single predictor variable transformations in the model and the non-transformation interaction terms. We then used stepwise AIC, stepwise BIC, and likelihood ratio tests on the new predictors to determine if the model needed further changes. The final model we created from stepwise AIC selection results was as follows:

$$\text{Log}(\pi/1-\pi) = -3.158 - 0.000004374*\text{Price} + 0.007901*\sqrt{\text{Price}} - 0.2907*(\text{AC=Yes}) + 0.1821*\text{Rooms} + 2.221*(\text{Rooms}^{0.2}) - 0.04816*\sqrt{\text{BEDRM}} + 0.1069*(\text{CNDTN=Excellent}) - 1.196*(\text{CNDTN=Fair}) + 0.2849*(\text{CNDTN=Good}) - 1.373*(\text{CNDTN=Poor}) + 0.6739(\text{CNDTN=Very Good}) - 0.4306*(\text{Ward=2}) - 0.2858*(\text{Ward=3}) - 0.0515*(\text{Ward=4}) + 0.2901(\text{Ward=5}) + 0.000001085\text{PRICE}*(\text{AC=Yes}) + 0.00000002.698*(\text{PRICE*ROOMS}) + 0.0000003.897(\text{Price*Ward=2}) + 0.0000005486*(\text{Price*Ward=3}) + 0.000001052*(\text{Price*Ward=4}) - 0.0000003.313*(\text{Price*Ward=5})$$

This model had an AIC of 16748. The interpretation of this model is as follows:

For every additional pricing dollar, the odds a property being qualified to sell decrease by a factor of 0.00000437402. For every additional square root of a pricing dollar, the odds a property is qualified to sell increase by a factor of 1.01. If a property has air conditioning, then the odds a property is qualified to sell increase by a factor of 0.34. For every additional room, the odds a property being qualified to sell decreases by a factor of 0.2. For every additional room raised to the one-fifth power, the odds a property being qualified to sell increases by a factor of 9.22. For every additional square root of a bedroom, the odds a property being qualified to sell decrease by a factor of 0.62. If a property has an excellent condition rating, then the odds that property being qualified to sell decreases by a factor of 2.31. If a property has a good condition rating, then the odds that property being qualified to sell increases by a factor of 1.33. If a property has a poor condition rating, then the odds that property being qualified to sell decreases by a factor of 2.95. If a property has a very good condition rating, then the odds that property being qualified to sell increase by a factor of 1.96. If a property is in Ward 2, then the odds that property being qualified to sell will decrease by a factor 0.54. If a property is in Ward 3, then the odds that property being qualified to sell will decrease by a factor 0.33. If a property is in Ward 4, then the odds that property being qualified to sell will decrease by a factor of 0.05. For every additional dollar added to the price and if the property has air conditioning, the odds a property being qualified to sell will increase by a factor of 1. For every additional dollar added to the price and for every additional room, the odds a property being qualified to sell will increase by a factor of 1. For every additional dollar added to the price and if the property is in ward 2, the odds a property being qualified to sell will increase by a factor of 1. For every additional dollar added to the price and if the property is in ward 3, the odds a property being qualified to sell will increase by a factor of 1. For every additional dollar added to the price and if the property is in ward 4, the odds a property being qualified to sell will increase by a factor of 1. For every additional dollar added to the price and if the property is in ward 5, the odds a property being qualified to sell will decrease by a factor of 0.0000003313001.

The reason we choose the stepwise AIC model was that we felt that the stepwise selection for BIC was penalizing our variables too much. With the creation of the final model, there was only the goodness of fit test left to run. Kingsley conducted this test and told me that the results were that the model fits the data. I do not believe this is accurate with such a high AIC

but one will have to trust Kingsley's judgment in this case. We also used the ROC Curve to determine well the training set model works compared with the validation set model. We can conclude that our model works well because the two areas were greater than seventy percent, even though we would have liked the areas to be above .85, and the training set and validation set lines were very similar, almost the same.

Conclusion

To conclude we have created a binary logistic model for what determines whether a property is qualified enough to sell. Although the AIC is very high as 16,748, it seems that the model fits that data well. It was a long analysis process but we could complete it even with the time restrictions we had. Now we will answer the 7 questions we had at the beginning of this study. The answers to our seven questions were as follows: 1) We were not able to hear back from Chris, the provider of this dataset on Kaggle, we so can still not answer what the qualification column in the original dataset means. 2) The qualifications for a residential property to be sold on the market is that the paperwork is completed and submitted, the bank approves any transaction that the buyers and sellers need and the inspection of the property is passed. 3) From all our analysis so far, we can conclude that property pricing is the most important factor in determining whether a property is qualified to go on the market 4) From all our analysis so far, we can conclude realtors do care whether the property is qualified to sell and money is the most important to them. Since the more the realtor sells and the higher the price, the property is sold for the more money from the deal they receive. 4) We believe that we were creating the most optimal regression for modeling properties based on our response variable being qualification. Overall though a multiple linear type regression would work best when it comes to making housing, property, and apartment models. 5) We did follow the previous housing model

approaches for predictor variables at the beginning of our model building but our analysis later had different predictor variables from those models creating using linear regression. 6) Yes, money is the most important thing. We will not say how this defines this world since we do not wish to be labeled as pessimistic people. Thankfully, we could answer our questions based on our analysis results and work, yet this does not mean we will stop the analysis being conducted.

For future analysis, we would do many things differently and add many different types of things. We will do the following things in the future: 1) conduct more time analysis and visualizations, 2) conduct some sentiment analysis on the street, neighborhood, and State one lives in since people are sometimes superstitious, 3) Try to see if we can hear back on what qualification meant in the dataset, 4) Add a few more variables – for example: Heat, and the interaction terms of heat and AC, 5) Collect data from realtor's websites and fill in the information ourselves since we were losing data constantly, 6) Add neighborhood rating, neighborhood review 7) Collect data from the surrounding states (West Virginia, Virginia, Maryland).

In conclusion, through all our analysis and graphs our model might have been able to measure the odds for determining qualifications. This model is nowhere near good enough to be published or presented in a conference. This was a great learning experience for careers even though the model we created is still inferior to a linear regression pricing model since we believe that money is the most important to realtors and people when it comes to the housing market. If you wish to know more about the data and analysis we have completed visit reference link 1. As a famous person once said, "failure is the mother of success."

Appendix

Table 1 – Variance Inflations Factor

PRICE	I(PRICE^0.5)	AC=Yes
103.585353	31.046266	3.078162
ROOMS	**ROOMS^0.2**	**BEDRM^0.5**
11.516670	9.250033	1.983626
CNDTN=Excellent	**CNDTN=Fair**	**CNDTN=Good**
1.162295	1.013751	1.377496
CNDTN=Poor	**CNDTN=Very Good**	**Ward 2**
1.002527	1.334604	5.393325
Ward 3	**Ward 4**	**Ward 5**
6.147429	4.980247	6.713264
PRICE*AC=Yes	**PRICE*ROOMS**	**PRICE*Ward 2**
20.203751	22.719759	14.479582
PRICE*Ward 3	**PRICE*Ward 4**	**PRICE*Ward 5**
5.167960	4.013284	4.430496

References:

1. https://aaronniecestro.shinyapps.io/DC-Housing/

2. McKay, Allie W. “Farmers' Markets vs. Food Deserts: Which Are Winning in DC?” *The Capital's Markets*, 31 July 2014, thecapitalsmarkets.wordpress.com/2014/07/31/farmers-markets-vs-food-deserts-which-is-winning-in-dc/.

3. Johnson, Matt. “Washington's Systemic Streets.” *Greater Greater Washington*, ggwash.org/view/2530/washingtons-systemic-streets.

4. “Money Is The Root Of All Evil Stock Photos and Images.” *Alamy*, www.alamy.com/stock-photo/money-is-the-root-of-all-evil.html.

5. “Types of Housing Models and Programs.” *The 519*, www.the519.org/education-training/lgbtq2s-youth-homelessness-in-canada/types-of-housing-models-and-programs.

6. Dobbins, Tim, and John Burke. “Predicting Housing Prices with Linear Regression Using Python, Pandas, and Statsmodels.” *Learn Data Science - Tutorials, Books, Courses, and More*, www.learndatasci.com/tutorials/predicting-housing-prices-linear-regression-using-python-pandas-statsmodels/.

7. Corsini, Kenneth Richard. “STATISTICAL ANALYSIS OF RESIDENTIAL HOUSING PRICES IN AN UP AND DOWN REAL ESTATE MARKET: A GENERAL FRAMEWORK AND STUDY OF COBB COUNTY, GA .” *A Thesis Presented to The Academic Faculty*, Georgia Institute of Technology, Dec. 2009, smartech.gatech.edu/bitstream/handle/1853/31763/Corsini_Kenneth_R_200912_mast.pdf .

8. "Regression Data for Inclusionary Housing Simulation Model | DataSF | City, and County of San Francisco." *San Francisco Data*, data.sfgov.org/Economy-and-Community/Regression-data-for-Inclusionary-Housing-Simulatio/vcwn-f2xk/data.

9. Leonard, Kimberlee. “What Forms Are Needed to Sell a Home by Owner?” *Home Guides | SF Gate*, 29 Dec. 2018, homeguides.sfgate.com/forms-needed-sell-home-owner-7271.html.

10. Leonard, Kimberlee. “What Is the Procedure for Closing a for Sale by Owner House Sale?” *Home Guides | SF Gate*, 15 Dec. 2018, homeguides.sfgate.com/procedure-closing-sale-owner-house-sale-65511.html.