Residential Properties in DC

Chirage Jhamb, Danlei Qian, Gaofeng Huang, Xi Zhang

MS, Data Science; The George Washington University, Washington DC 20052; All contributors conducted equal research and analysis for this project

Introduction

Human beings spend most of our long lives at home. Therefore, houses are of great significance to us. Since in ancient times when people lived in caves, until the modern era. We are always chasing the best living conditions we can. Fortunately, we have learned many scientific and effective data tools to help us analyze and select people's dream houses in DC. We have shown the effect of different conditions on price in the last project. I believe that some buyers can choose the ideal house according to the current economic conditions. In this project, it is based on the first one, promotes some methods and focus on prediction.

In the first project, we mainly use linear regression models to find the impact of Floor Plan, Location, Time and Facilities on prices. Then we use the ANOVA test and correlation plot to confirm our assumption. However, due to the limitations of the linear regression model and the complexity of data. Our final result was not exact. After continuing learning, we have mastered more methods to classify and deal with different types of variables, and we believe that we can get more accurate and satisfied result. Furthermore, the methods we learned, such as KNN and Time series, enable us to make more accurate predictions with existing data.

Based on the progress of the previous project and what has been learned. We have three SMART Questions about today's topic:

1. *Whether the price and sales volume have some systematic pattern over the time period?*
2. *How to predict the price by K nearest neighbor?*
3. *How to predict and classify the condition of a property?*

The dataset was derived from Kaggle, which was derived from *https://www.kaggle.com/christophercorrea/dc-residential-properties*. We have completed data cleaning in the first project. The file DC_Properties.csv contains 125k rows and 49 columns. The selected columns are as follows:

1) BATHRM - Number of Full Bathrooms
2) HF_BATHRM - Number of Half Bathrooms (no bathtub or shower)
3) HEAT - Heating
4) AC - Cooling
5) NUM_UNITS - Number of Units
6) ROOMS - Number of Rooms
7) BEDRM - Number of Bedrooms
8) AYB - The earliest time the main portion of the building was built
9) YR_RMDL - Year structure was remodeled
10) EYB - The year an improvement was built more recent than actual year built
11) STORIES - Number of stories in primary dwelling
12) SALEDATE - Date of most recent sale
13) PRICE - Price of most recent sale
14) QUALIFIED - Qualified
15) SALE_NUM - Sale Number

16) GBA - Gross building area in square feet
17) BLDG_NUM - Building Number on Property
18) STYLE - Style
19) STRUCT - Structure
20) GRADE - Grade
21) CNDTN - Condition
22) EXTWALL - Exterior wall
23) ROOF - Roof type
24) INTWALL - Interior wall
25) KITCHEN - Number of kitchens
26) FIREPLACES - Number of fireplaces
27) LANDARE - Land area of property in square feet
28) WARD - Ward (District is divided into eight wards, each with approximately 75,000 residents)

Materials and Methods

The DC_Properties.csv was imported into RStudio Version 1.1.463, for analysis (RStudio, 2009- 2018). We use the following R Packages and libraries: library("bestglm"), library ("glmnet"), library ("gmodels"), library ("pROC"), library ("pscl"), library ("rpart"), library ("rpart.plot"), library ("randomForest"), library ("C50"), library ("rpart"), library ("rpart"). The cleaned data we use for our SMART questions will vary slightly. We will show our analysis and prediction in different aspects:

Whether the price and sales volume have some systematic pattern over the time period?

In this section, we would like to use time series analysis to explore some patterns of the properties' average monthly price and average monthly sales volume in D.C. from the history data and make a prediction.

Time series on price

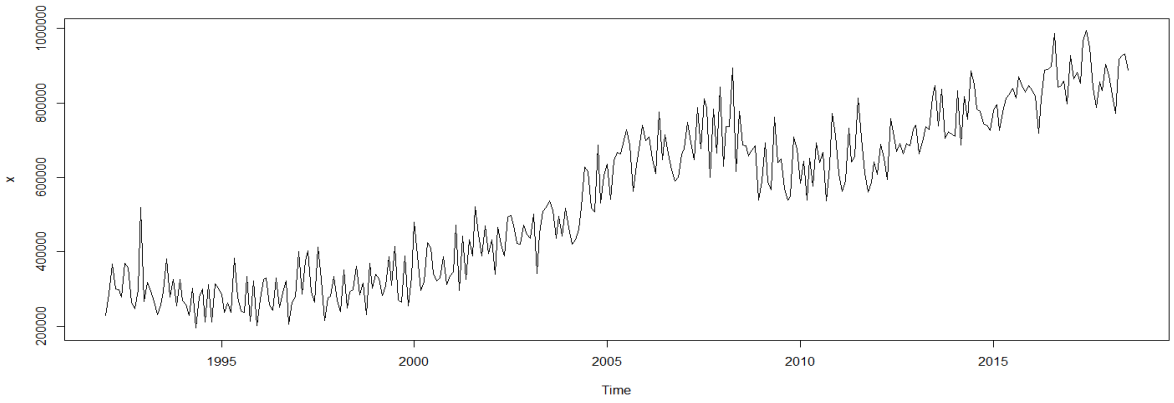


Figure 1 Create time-series objects for price from 1992

We selected "SALTEDATE " and" PRICE ", and used aggregate() function to calculate the mean value of properties' price per month from the past years as the variables.

The first process is to create the time-series objects and decompose the data to the four level including “random”, “seasonal”, “trend” and “observed”. There are many fluctuations and an

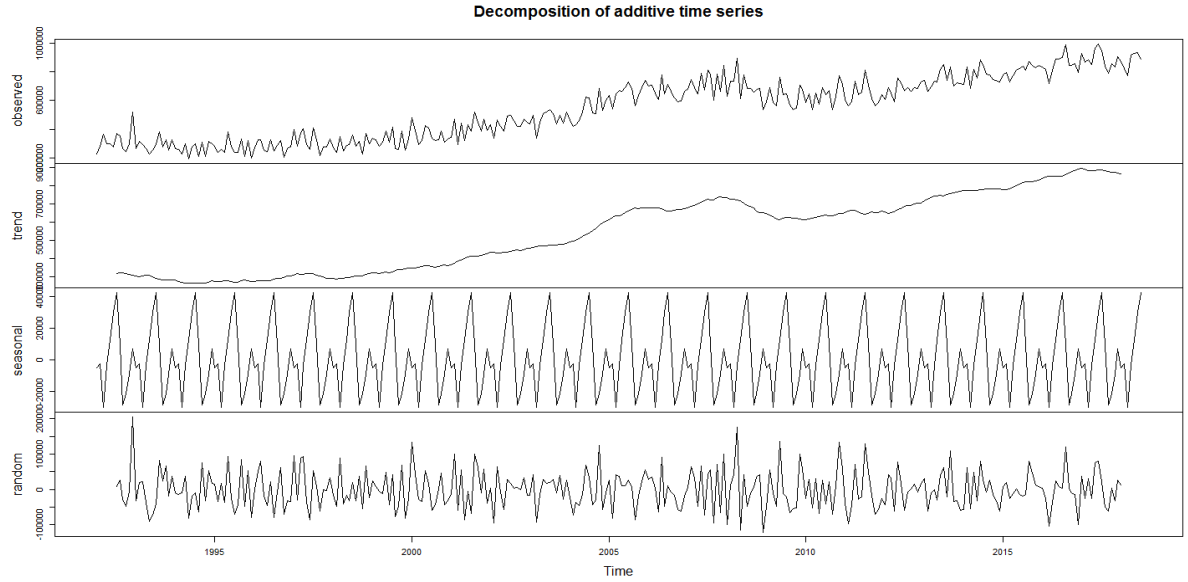


Figure 2 Decomposition of the price from 1992

overall increasing trend. We can see a sharp drop around 2008, and we may assume that this decrease was resulted from the terrible financial crisis. Also, there seems to be some seasonality, but not clear.

Then after using the Holter Winter function to smooth the data, we found the alpha, beta and gamma are closer to 0, which means the value is more based on the historical data but not recent data. And the SSE is 1414128761359, which is too high. At the same time, the forecast result also shows the inconsistent. In other words, this model cannot show a good fit for the price. Thus, it can be concluded that its hard to find a time pattern on price.

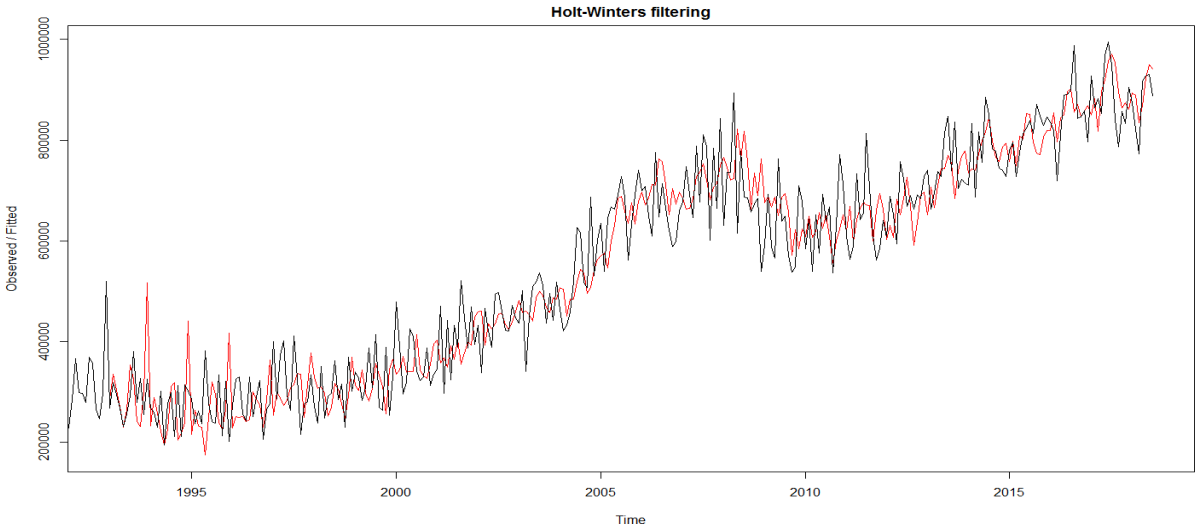


Figure 3 Forecast plot of price from 1992(black: original values; red: predicted values)

Time series on sales *volume*

Turning now to the research on the sales volume, we selected “SALEDATE” to calculate the sales volume per month and conducted the same procedure. As the forecast result shows that

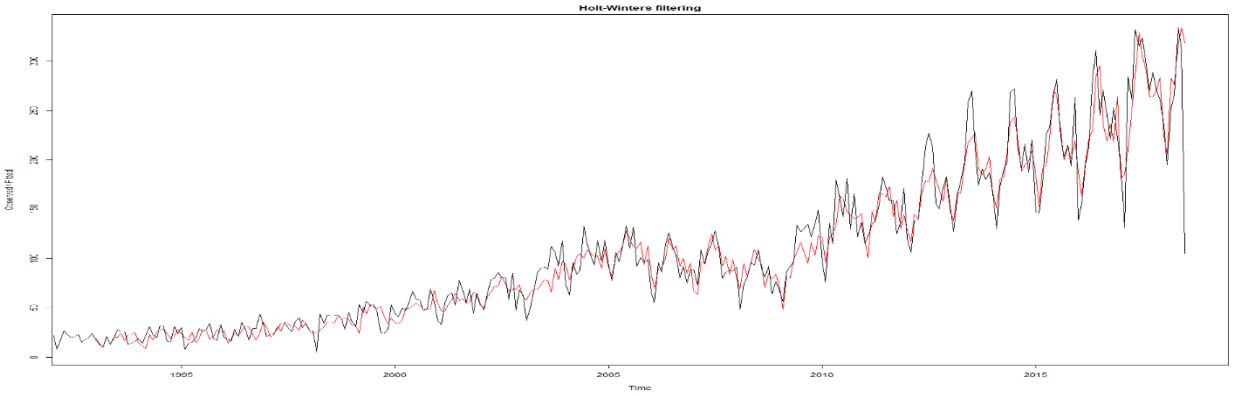


Figure 4 Forecast plot of sales volume from 1992(black: original values; red: predicted values)

even though the overall SSE is still high (132759.8), we can find the forecast for recent years seems fit well. Therefore, we move our object to the pattern on recent years. And we chose the range of time as 2013 to 2018.

After the time-series object of sales volume from 2013 to 2018 was set, it is clear that there is some seasonality. We can see the sales volume is always small at the start of a year and large at the middle of a year, as the adjusted seasonal component shows. After we use the "Holter

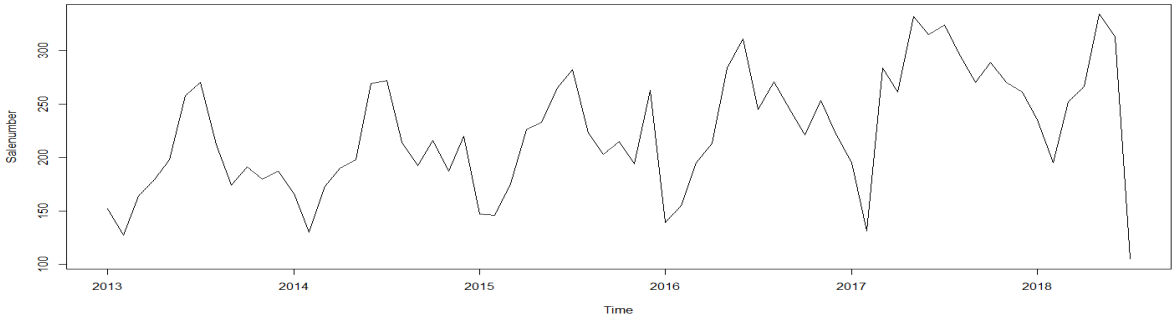


Figure 5 create time-series objects for sales volume from 2013

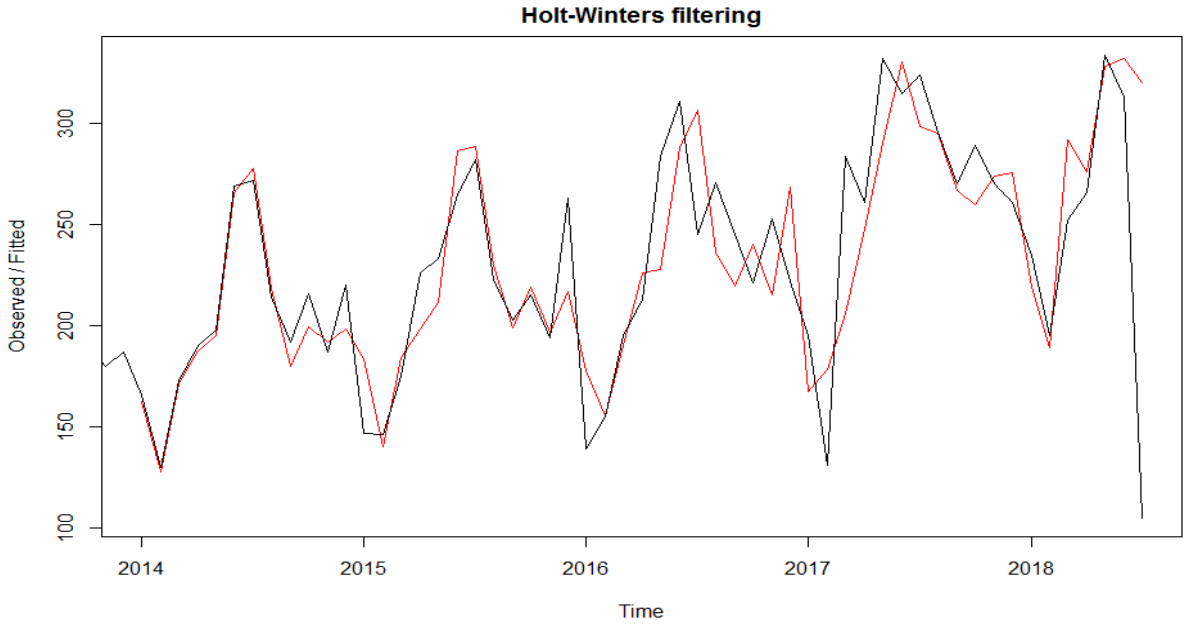


Figure 6 Forecast plot of sales volume from 2013(black: original values; red: predicted values)

Winter ()" function, the parameters look better. To be specific, "alpha=0.2" means the influence weight of recent observations is small; "beta=0" means the slope of the trend remains constant throughout the whole time series; "gamma=0.64" means that seasonal partial predictions are based on both the recent observations (more) and history observations. Also, it can be seen that the forecast plot appears to fit well since the forecast value is more consistent with the original observations. Besides, the SSE become smaller (83028.09). At last, we made a prediction for the future 12 months based on the results. We can see that the next peak value is predicted to be in the middle of 2019 while the valley value is predicted to be at the beginning of 2019. Also, there will be a slump after the peak.

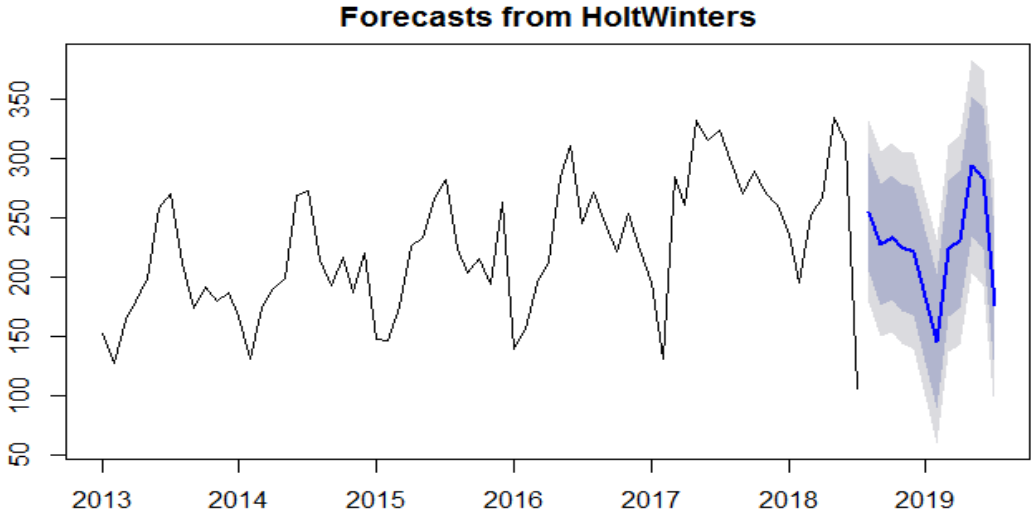


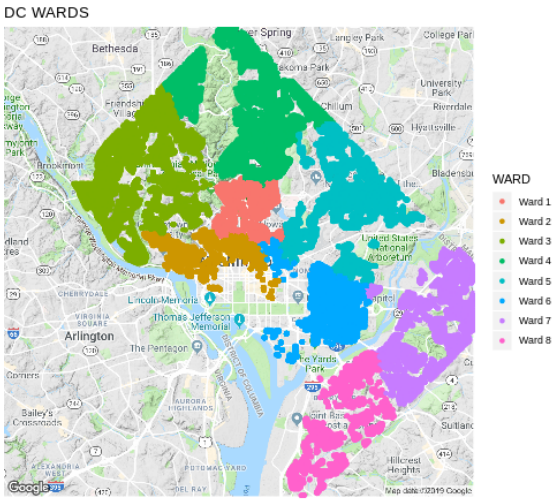
Figure 7 Future prediction of sales volume for 12 months

Limitation

Since the case for price is not a good fit for time series analysis, we can learn that stationarity is quite important in the time series. However, we lack the process of stationarity tests to check if the data is stationary. Besides, only Holt-Winters' additive method was used in this study. More methods need to be explored to enhance the fit of time series theory.

How to predict the price by K nearest neighbor?

For K nearest neighbor, we need to first put the variable we're predicting into categorical form. Due to this limitation, we use Quantiles. Initially, for predicting the Quantile of the price we chose "LATITUDE", "LANDAREA","YR_RMDL","GRADE", "WARD", "ASSESSMENT_NBHD". The accuracy was low, even after multiple attempts at eliminating columns with low relevance. Then, distributing the dataset into different sections by using the "WARD" variable made some improvement. In each "WARD", selling price is based upon different unique factors which might not be of relevance in other wards. Individually, the accuracy score of each ward using KNN was 0.64 to 0.78, as opposed to the previous accuracy score of 0.48. Also, by using decision tree for each ward we eliminated certain columns during each classification of KNN. These were the columns which showed no importance for that particular ward. Then while creating the final model, only the most common columns were chosen. These gave the final accuracy score of 0.59.



How to predict and classify the condition of a property?

Apart from the price, the condition of a property is always a determinant factor we will concern when we plan to buy a property. In this section, our purpose is going to predict the condition of a particular property without seeing the photos of it. Here, we can hardly get the specific features of a property that show the condition of this property. In other words, we cannot find the causality between some features and the condition of a property. Thus, in this report, we just try to predict a property tends to be in better condition through some general features which can be easily collected.

1) Definition of condition.

Analogous to the problem in the ranking method (e.g. customer reviews in Amazon), the assessment of the condition is varied from different individuals. Hence, there is no uniform rule to classify properties into different levels of condition. Here is a detailed explanation of the condition from Marshall & Swift Condition Assessment (page E-6).

Excellent Condition - All items that can normally be repaired or refinished have recently been corrected, such as new roofing, paint, furnace overhaul, state of the art components, etc. With no functional inadequacies of any consequence and all major short-lived components in like-new condition, the overall effective age has been substantially reduced upon complete revitalization of the structure regardless of the actual chronological age.

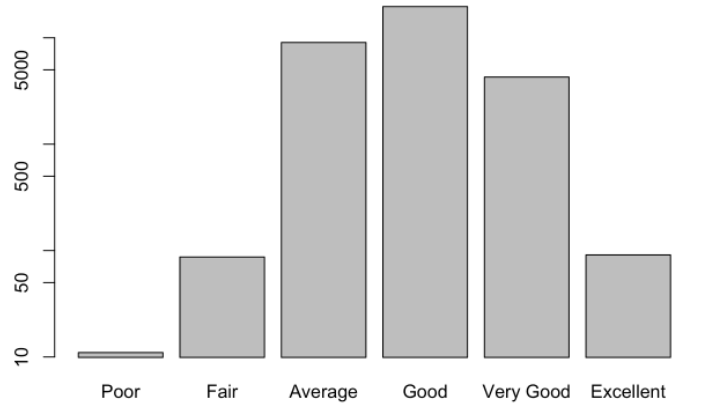
Very Good Condition - All items well maintained, many having been overhauled and repaired as they've shown signs of wear, increasing the life expectancy and lowering the effective age with little deterioration or obsolescence evident with a high degree of utility.

Good Condition - No obvious maintenance required but neither is everything new. Appearance and utility are above the standard and the overall effective age will be lower than the typical property.

Average Condition - Some evidence of deferred maintenance and normal obsolescence with age in that a few minor repairs are needed along with some refinishing. But with all major

components still functional and contributing toward an extended life expectancy, effective age and utility are standard for like properties of its class and usage.
Fair Condition (Badly worn) - Many repairs needed. Many items need refinishing or overhauling, deferred maintenance obvious, inadequate building utility and services all shortening the life expectancy and increasing the effective age.
Poor Condition (Worn Out) - Repair and overall needed on painted surfaces, roofing, plumbing, heating, numerous functional inadequacies, substandard utilities etc. (found only in extraordinary circumstances). Excessive deferred maintenance and abuse, limited value-in-use, approaching abandonment or major reconstruction, reuse or change in occupancy is imminent. Effective age is near the end of the scale regardless of the actual chronological age.

```
     Poor      Fair   Average      Good Very Good Excellent
       11        87      9038     19661      4277        91
```



2.) A glance at the condition

From the distribution of the number of properties with respect to different conditions, it looks like a lanky normal distribution, which is reasonable. Over 99% of properties are in “Average”, “Good”, and “Very Good” condition. Therefore, it may cause some problems (will discuss later) to predict the condition of the other three levels, which is “Poor”, “Fair”, and “Excellent.”

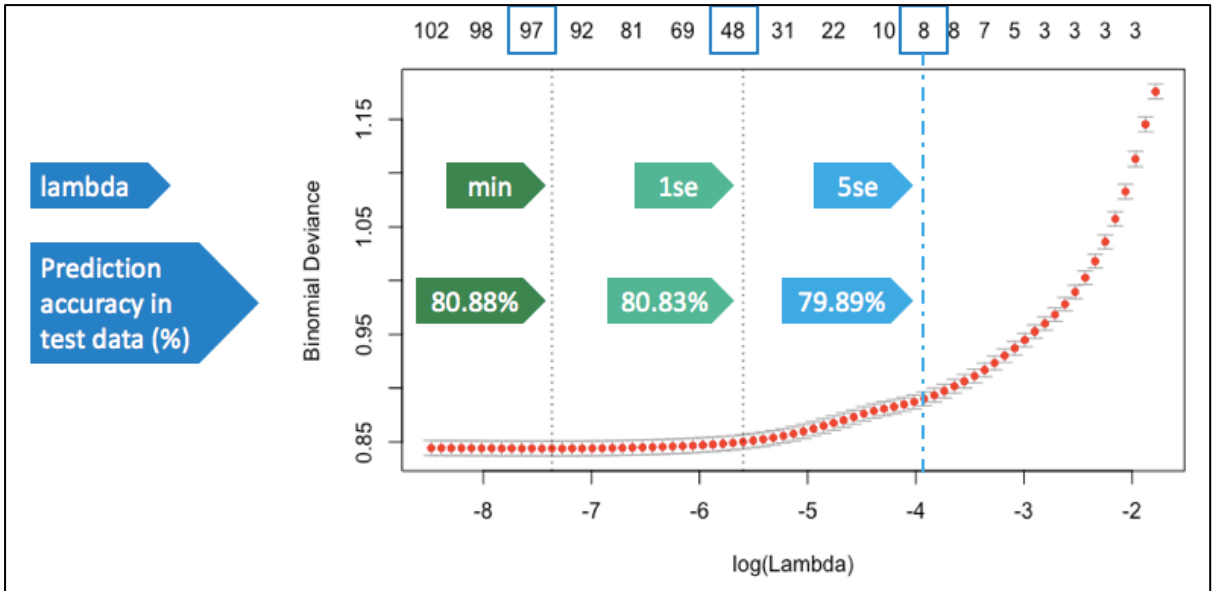
3) Binomial prediction [with simplification]

For simplicity, we first try to distinguish whether a property is above or below average condition. In other words, we trivially split the condition into two levels, “<= Average” (including “Poor”, “Fair”, “Average”) and “> Average” (including “Good”, “Very Good”, “Excellent”).

```
<= Average  > Average
      9136       24029
```

4) LASSO-logistic regression [feature selection]

As the condition grouped into 2 levels, we can apply the logistic regression to solve this binomial prediction problem. However, unfortunately, it does not select out a small group of variables when the lambda is within 1 standard error (over 48 features).



Thus, we try to choose the model with 8 features in 5 standard error away from the best model. Although selecting a model outside 1 standard error will lead to somewhat bias, it performs well in the prediction of test data. As the prediction accuracy in the best model is 80.88%, this simplified model has a pretty good accuracy of 79.89%.

	observed.classes		
predicted.classes	0	1	Row Total
0	1297	446	1743
	0.744	0.256	0.155
	0.416	0.055	
	0.115	0.040	
1	1822	7712	9534
	0.191	0.809	0.845
	0.584	0.945	
	0.162	0.684	
Column Total	3119	8158	11277
	0.277	0.723	

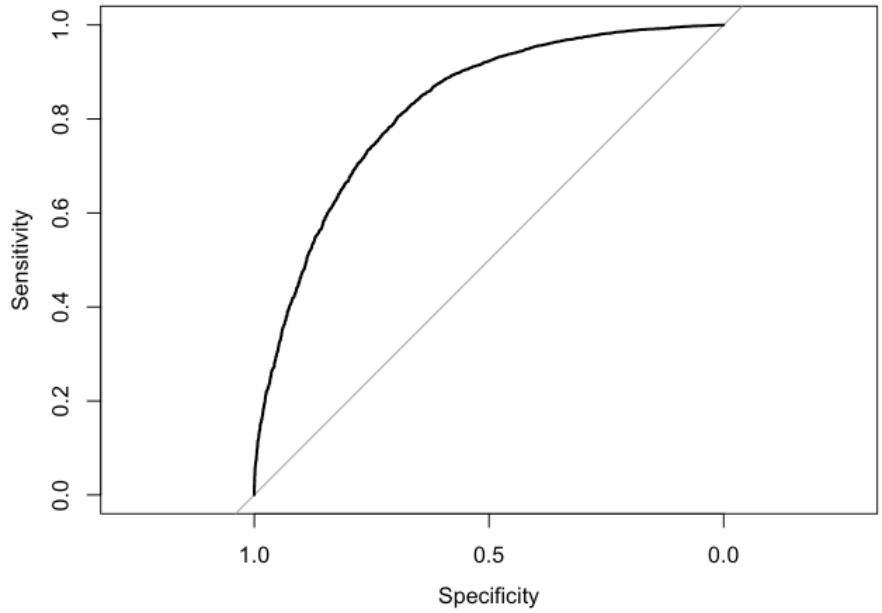
As for the confusion matrix of the prediction result, this model performs better in predicting a property with the above average condition. Here, in the test data, the accuracy is 80.9% when a property is predicted as the above average condition. Besides, 94.5% of above-average properties are correctly predicted in the test data.

Generalized linear model and evaluation

LASSO selects out 8 variables (“HEAT”, “AC”, “AYB”, “YR_RMDL”, “EYB”, “PRICE”, “QUALIFIED”, “SALE_NUM”), then we take the 2nd-round feature selection through the best subset GLM. Then, “HEAT” is also moved out. In fact, “HEAT” does not contribute so much in the prediction.

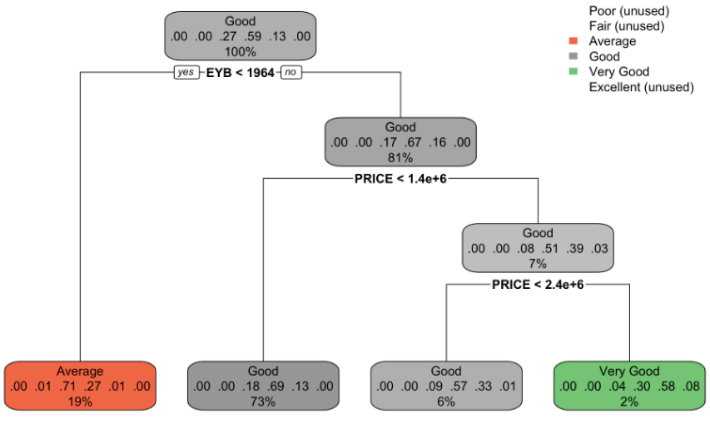
After the feature selection, we get a simple GLM to predict the condition (above or below average). This model performs quite well in ROC curve where AUC is greater than 0.8.

```
## Area under the curve: 0.8234
```



As for the McFadden's pseudo R^2 value, it has different evaluation criteria comparing with the R^2 value. McFadden states "while the R2 index is a more familiar concept to the planner who is experienced in OLS, it is not as well behaved as the rho-squared measure, for ML estimation. Those unfamiliar with rho-squared should be forewarned that its values tend to be considerably lower than those of the R2 index…For example, values of 0.2 to 0.4 for rho-squared represent EXCELLENT fit." If we get such value of a GLM from 0.2 to 0.4, it indicates this model can explain most of the data.

5) Prediction of 6-level condition [without simplification]




```
## [1] 0.67314
```

We've solved the two-level prediction. Now we keep attacking the 6-level prediction. Using classification tree is a better way to predict a categorical variable with multi-levels.

5.1) (Single) Decision tree

We apply the selected variables in the previous feature selection result to this decision tree. Here, the tree only uses two variables, "EYB" and "PRICE." From this simple prediction model, it shows an improvement after 1964 and the price of over 2.4 million means the property tends to be better. However, there are two defects of this model. First, the total prediction accuracy is only 67.3%. Also, in the confusion matrix, it performs not well in the prediction of "Very Good" condition, which is

only 51.1%. Second, this model misses predicting 3 minor levels, which are “Poor”, “Fair”, and “Excellent”.

5.1) Random Forest

To improve the prediction accuracy of the single tree model, we try to apply the random forest algorithm. The prediction accuracy works better in this improved model, but it still misses 3 minor levels.

6) Limitation

Because the criteria of the decision tree are information gain (Shannon Entropy), it tends to make a decision close to the majority. Here, most of the data are the condition of good, very good, and average. Thus, this tree will not tend to make a decision on the condition of excellent, poor, and fair.
We have tried to solve this “discrimination” problem, but still in vain. Therefore, we need a model can equally treat all levels in a categorical variable although the training data for each level is not in the same size.

Conclusion

Through our research and prediction, we did a more in-depth analysis, make some predictions and find out the limitation. At the same time, we hope to give some advice to people who want to buy a house in DC:

1) Price & Sales Volume
   a) Analysis: Prices continue to rise with irregular fluctuations, while sales have fluctuated periodically in recent years. We find that sales are more in summer than in winter.
   b) Prediction: Since the overall trend is also increasing, we expect property prices and sales to continue to rise overall over the next 12 months, so it's wise for you to buy early.
   c) Limitation: While the overall trend can be shown clearly, short-term prices are not easily predicted by time series.

2) Location
   a) Analysis: We used the decision tree to make feature selection for different wards, then we find out the most important variables for all of the wards.
   b) Prediction: After feature selection, the final KNN model can predict the price by the nearest neighbors.
   c) Limitation: Because of the limitations of KNN, we can only predict the range of prices but not accurate numbers.

3) Quality
   a) Analysis: Since we have mastered more methods of dimensionality reduction (such as LASSO, GLM, etc.), we can more accurately select the variables we need. In terms of condition prediction, the result of differ good or poor based on average is more accurate.
   b) Prediction: If you want to buy a house of above-average quality, you only need to pay attention to whether it has air conditioning, the year it was built, the year it was remodeled, the year it was improved, the price, whether it is qualified, and the number of sales.

c) Limitation: However, the existing model (decision tree and random forest) still has some deficiencies due to more accurate classification.

Reference

ChrisC (2018). D.C. Residential Properties: Residential Properties in Washington, from https://www.kaggle.com/christophercorrea/dc-residential-properties

Marshall & Swift Condition Assessment (page E-6).