Sean Heffernan

Thilanka Munasinghe

Data Analytics

November 4, 2022

Data Analytics Assignment 4

* 1. To conduct an effective data analysis, I should be looking at trends surrounding the sales of buildings in Manhattan by looking at various factors which might influence final sales price. This includes features like the zip code or latitude/longitude (location may indicate higher or lower-end neighborhood), size of the property (in square feet) and building class. To that end, I will be investigating these features individually first, to see if I can come to any conclusions surrounding them. For example, to make these simple I will be looking at the sizes of the properties and the building class to see if there are any trends surrounding purchases made - are more buildings office buildings? Store buildings? Etc. How large are the purchase sizes (square feet-wise), and does that influence the total cost? I used both histograms and boxplots to visualize these data points as well as the summary method to view specific stats

A picture containing box and whisker chart

Description automatically generated

> summary(manhattan\_sales$YEAR.BUILT)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0 1912 1940 1745 1975 2019 1871

The average year of building being built was 1745. Interestingly, some data points have NA as a year built while others have 0 - maybe the pre-processing for this data set was inconsistent?

Chart

Description automatically generated

> summary(manhattan\_sales$SALE.PRICE)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000e+00 8.324e+04 8.000e+05 3.018e+06 1.860e+06 2.398e+09

The most expensive property was a whopping 2 billion dollars on 75 9 Avenue. Doing some outside research, it appears building was bought by Google for office space.

Chart

Description automatically generated with medium confidence

> summary(manhattan\_sales$SALE.PRICE)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000e+00 8.324e+04 8.000e+05 3.018e+06 1.860e+06 2.398e+09

IQR has been calculated by getting the 1st and 3rd quartiles via the summary() method and getting the difference between the two.   
1860000 - 83240 = 1776760.

Lower Inner Fence: 83240 - 1.5\*1776760 = -2581900

Upper Inner Fence: 1860000 + 1.5\*1776760 = 4525140

Outliers found by getting values outside the lower and upper inner fences. There are 7312 outliers.

* 1. (More details about how I conducted my Multivariate Regression are in the code file.) MSE of each result were extraordinarily high and the multiple r squared values were extremely low (MSE was in the quadrillions lol and R squared was below 0.4 indicating low correlation). What this tells me is that the Land and Gross square feet of a property aren't nearly enough to predict what the potential price of a property will be. This comes to no surprise considering the different locations of properties around Manhattan areas with higher cost of living compared to others. For each sample of data I took roughly half of the dataset used it as training data. While there were some variations on the R squared value as well as the MSE, they remain consistently around the same values.
  2. Multiple steps needed to be taken in order to clean the data (many of which I did in 1c also). For starters, I needed to be conscious of data which \*appeared\* to be legible but in fact were being read incorrectly by R. For example, the Land and Gross square feet of properties were strings and not numerics, so couldn't be properly used as training features until being parsed out as numerics. Additionally, many of the columns have NA values or just 0 as a value, which can lead to an inaccurate model. For that sake, it is necessary to clean out and remove those rows if the feature I wish to look at has 0 or NA. Finally, it would be good to remove the outliers for the sale price, as these could negatively influence the model. As far as model fits go, Multivariate regression w/ Land and Gross square feet is not enough to create a fitting model for this data as proven in the previous question. Due to the accurate nature of Random Forest due to the many different regressions used, I will use that to model the same data as before and see if the same conclusions are met.
  3. Conducted random forest analysis as described in the previous question:

> importance(model)

%IncMSE IncNodePurity

LAND.SQUARE.FEET 7.363157 1.155277e+19

GROSS.SQUARE.FEET 26.425115 2.023425e+19

Diagram

Description automatically generated

After printing out the mean squared error (not as high as multivariate regression but still very high in the hundred billions), we can see that using land square feet and gross square feet alone is still inaccurate to properly predict sale price. However, the random forest helps us determine that the gross square feet is more important compared to land square feet as far as an increase in MSE. By using varImpPlot() it is clear that the removal of GROSS SQUARE FEET leads to the largest percentage in increased MSE (around 26%) while the removal of LAND SQUARE FEET in the prediction only increases MSE by 7.36%

> model

Call:

randomForest(formula = SALE.PRICE ~ LAND.SQUARE.FEET + GROSS.SQUARE.FEET, data = manhattan\_clean, importance = TRUE)

Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 1

Mean of squared residuals: 1.671538e+15

% Var explained: 29.74

As accomplished in the previous step, we can verify the fit of the model by looking at the MSE of the validation set. I calculated the MSE to be in the hundred billions (6.051824e+14 to be exact in my case), which is extremely high, indicating that these two variables are not enough to accurately predict SALE.PRICE. The r squared value supports this theory, with an R squared value of 29.74 (the % Var explained value). Overall, the MSE is much lower compared to the multivariate regression I performed in 1c, but the previous discovery that GROSS and LAND SQUARE FEET are not enough to properly predict sales price remains true.

* 1. Overall I found it frustrating that the types of variables for each feature varied even when they probably shouldn't have (the prime example here is the strings for land and gross square feet vs the already numeric prices). Without proper discovery beforehand it can lead to hours wasted trying to debug when models when perhaps it's the data which was inaccurate. This just goes to show how important preprocessing and data cleaning is when it comes to large datasets. When first conducting the multivariate regression, I was very unsure about the results I was getting due to the obscenely high MSE values and low R squared values. However, after doing some more thinking (it is unlikely for two variables in a 30 column dataset to accurately predict EVERYTHING) and the follow-up random forest model, I was more confident with my conclusions.

1. Of the two model types, random forest seems to outperform multivariate regression in this case. This isn't that surprising considering one of random forest's strengths being good to create estimates for missing data (in our case we only used 2 features which made up for a small fraction of the data). Before resorting to random forest, I had also tried KMeans clustering for 2a to try and make predictions for sale price. However, there is no native predict() method that works with KMeans, so I couldn't use it for regression as I had intended for. That being said, plotting out the KMeans was still useful for visualizing the relationship between price and gross square feet and identifying some clustering - but again, it couldn't exactly be used to predict the desired label (predicting prices for certain gross & land square feet). As an overall, it's apparent that a fine balance needed to be struck between too many and too little features being used when creating an accurate regression model in efficient time. Too many features and the model creation might take too long, too few features and you risk an inaccurate model (like the one we put together during this assignment).