Yahya Alhinai

EE5373

November 26th, 2019

Data Segmentation for House Price Predictions Project

OBJECTIVE FOR THIS LAB:

- This lab introduces you to data segmentation and the Kaggle competition process

1. Development of the "all predictors" model is explained, and the model is shown:

The predictors that were removed are:

$Id, prop\_id, date, waterfront$

because those predictors have no contribution on the price whatsoever.

The prediction that was used are the rest which:

$$bedrooms, bathrooms, sqft\_living, sqft\_lot, floors, view, condition, grade, sqft\_above,$$
$$sqft\_basement, yr\_built, yr\_renovated, zipcode, lat, long, sqft\_living15, sqft\_lot15$$

The following predictions coefficient is generated through RStudio:

```
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    7.313e+06  3.544e+06   2.064  0.03907 *  
bedrooms      -4.086e+04  2.245e+03 -18.199  < 2e-16 ***
bathrooms      4.006e+04  3.914e+03  10.235  < 2e-16 ***
sqft_living    1.637e+02  5.274e+00  31.050  < 2e-16 ***
sqft_lot       3.473e-02  6.239e-02   0.557  0.57782    
floors         3.650e+03  4.319e+03   0.845  0.39797    
view           7.897e+04  2.372e+03  33.296  < 2e-16 ***
condition      2.919e+04  2.845e+03  10.259  < 2e-16 ***
grade          9.194e+04  2.585e+03  35.560  < 2e-16 ***
sqft_above     3.859e+01  5.236e+00   7.371 1.77e-13 ***
sqft_basement         NA         NA      NA       NA    
yr_built      -2.552e+03  8.748e+01 -29.173  < 2e-16 ***
yr_renovated   2.610e+01  4.404e+00   5.926 3.17e-09 ***
zipcode       -6.051e+02  3.978e+01 -15.209  < 2e-16 ***
lat            6.023e+05  1.297e+04  46.446  < 2e-16 ***
long          -2.271e+05  1.598e+04 -14.212  < 2e-16 ***
sqft_living15  8.677e+00  4.148e+00   2.092  0.03645 *  
sqft_lot15    -2.636e-01  8.972e-02  -2.937  0.00331 ** 
```

This model predicts the house prices with 95% confidence interval:

$$95\% \text{ confidence interval} = 111.87 \pm 210.41$$

```
        One Sample t-test
data:  new
t = 1.042, df = 80999, p-value = 0.2974
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -98.54913 322.29032
sample estimates:
mean of x
 111.8706
```

The predictions of the price of houses are not to be considered as a reliable source because the prices given are not accurately produced.

2. Development of the full backward elimination model is explained and the model is shown:

We started having all the predictors selected be evaluated for the regression model. We obtained the following:

$$R^2: 0.2346 \,||\, Adjusted - R^2: 0.2269$$

sqft_basement: seems to have a lot of NA values. It is to be removed

$$R^2: 0.2346 \,||\, Adjusted - R^2: 0.2269$$

long has a high p-value. The following R-value after removing both:

$$R^2: 0.2346 \,||\, Adjusted - R^2: 0.2274$$

sqft_lot15 has a high p-value indicating low contribution. The following R-value after removing both:

$$R^2: 0.2345|| \, Adjusted - R^2: 0.2278$$

sqft_above shows low contribution to the model. It is to be removed:

$$R^2: 0.0.2342 \,||\, Adjusted - R^2: 0.228$$

sqft_living15 add no significant values to the price predicting model:

$$R^2: 0.2338 \,||\, Adjusted - R^2: 0.2281$$

zipcode has no to little effect and therefore it is to be removed:

$$R^2: 0.2332|| \, Adjusted - R^2: 0.2279$$

The **sqft_lot** predictor has a p-value that is 0.15 which is out of the desirable accuracy:

$$R^2\text{: } 0.2314 \;||\; Adjusted - R^2\text{: } 0.2271$$

bedrooms has a high p-value the is not effecting the model, it's to be removed:

$$R^2\text{: } 0.2303 \;||\; Adjusted - R^2\text{: } 0.2265$$

The following predictions coefficient is generated through RStudio after the elimination process:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.860e+06  3.968e+05  -7.207 8.73e-13 ***
bathrooms     7.236e+03  2.121e+03   3.412 0.000662 ***
sqft_living   2.372e+01  2.876e+00   8.247 3.33e-16 ***
view          1.478e+04  5.152e+03   2.869 0.004166 **
condition     1.006e+04  1.147e+03   8.776  < 2e-16 ***
grade         8.800e+03  1.322e+03   6.657 3.81e-11 ***
yr_built      1.215e+02  4.562e+01   2.662 0.007839 **
yr_renovated  5.311e+00  2.362e+00   2.248 0.024685 *
lat           5.680e+04  7.664e+03   7.410 2.02e-13 ***
```

This model predicts the house prices with 95% confidence interval:

$$95\% \text{ confidence interval} = -24.02 \pm\ 210.9$$

```
       One Sample t-test
data:  new
t = -0.2233, df = 80999, p-value = 0.8233
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -234.9339  186.8779
sample estimates:
mean of x
-24.02802
```

Although the result generated after going through elimination prosses are better that the previous version as the mean is closer to zero, the rage is still out of the range of our desire. The confidence interval is wider than it should be, and it is still not very close to the value zero. This is due to the multiple factors that effecting the price of a house and the how much each factor is effecting the price which make it hard to predict each house price accurately.

3. Development of the zip code segmentation models are explained and the models are shown.

This section is meant to divide housing prices into 3 categories based on the average prices for each zip code. This is done in favor of getting a better regression model to accurately describing houses' price. The three subsets are zip code with the highest average house price, zipCode with the lowest average house price, and zip code with the median average house price.

We found the following:

$highest\ average\ house\ price\ zip\ code\ =\ \mathbf{98002}$

$median\ average\ house\ price\ zip\ code\ =\ \mathbf{98059}\ and\ \mathbf{98011}$

$lowest\ average\ house\ price\ zip\ code\ =\ \mathbf{98039}$

Each of these categories were separated in order to generate a regression model for each set after going though backward elimination. Compare each set with the other two and determine how good the model describing each set. 95% confidence interval were used to obtain the range from the mean. For each set it was halved into two subsets for train set and for test set.

A. Highest Average House Price:

After going through the elimination process the predictions used are:

$$lm(formula\ =\ price \sim sqft\_living\ +\ view\ +\ condition\ +\ grade\ +\ sqft\_above)$$

With

$$R^2\text{: }0.9374||\ Adjusted - R^2\text{: }0.9272$$

The coefficients for the predictions are

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.860e+06  7.410e+05  -2.510  0.01752 *  
sqft_living  4.357e+02  8.154e+01   5.344 8.02e-06 ***
view         1.655e+05  5.977e+04   2.770  0.00939 ** 
condition    2.142e+05  1.205e+05   1.777  0.08534 .  
grade        1.102e+05  7.626e+04   1.445  0.15860    
sqft_above   1.424e+02  9.862e+01   1.444  0.15889    
```

B. Median Average House Price:

After going through the elimination process the predictions used are:

$$lm(formula\ =\ price \sim bedrooms\ +\ sqft\_living\ +\ sqft\_lot\ +\ floors\ +$$
$$condition\ +\ grade\ +\ sqft\_above\ +\ zipcode\ +\ lat\ +\ sqft\_living15)$$

With

$$R^2: 0.8496 || Adjusted - R^2: 0.8465$$

The coefficients for the predictions are

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -7.972e+08  1.424e+08  -5.597 3.69e-08 ***
bedrooms      -1.638e+04  5.771e+03  -2.838  0.00474 **
sqft_living    6.816e+01  1.284e+01   5.308 1.70e-07 ***
sqft_lot       6.573e-01  1.644e-01   3.999 7.36e-05 ***
floors        -5.063e+04  1.190e+04  -4.256 2.51e-05 ***
condition      1.974e+04  7.366e+03   2.680  0.00762 **
grade          6.569e+04  5.441e+03  12.072  < 2e-16 ***
sqft_above     3.938e+01  1.307e+01   3.013  0.00273 **
zipcode        7.406e+03  1.337e+03   5.541 5.00e-08 ***
lat            1.487e+06  2.416e+05   6.156 1.59e-09 ***
sqft_living15  7.152e+01  9.495e+00   7.532 2.54e-13 ***
```

C. lowest Average House Price:

After going through the elimination process the predictions used are:

$$lm(formula = price \sim bathrooms + sqft\_living + sqft\_lot + condition + yr\_built + lat)$$

With

$$R^2: 0.7333 || Adjusted - R^2: 0.7222$$

The coefficients for the predictions are

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.435e+07  6.032e+06  -2.378  0.01871 *
bathrooms    1.147e+04  5.568e+03   2.060  0.04120 *
sqft_living  5.718e+01  7.302e+00   7.830 9.68e-13 ***
sqft_lot     2.224e+00  7.506e-01   2.962  0.00357 **
condition    1.156e+04  5.023e+03   2.302  0.02280 *
yr_built     6.042e+02  1.196e+02   5.052 1.30e-06 ***
lat          2.794e+05  1.282e+05   2.180  0.03091 *
```

The following table is meant to compare how the three regression models holds when used to describe different sets:

ZipCode	Lowest average	Median average	Highest average
Lowest average	-46.54565 ± 628.01695	-114695.8 ± 2034.75	-1801142 ± 50083
Median average	-634570 ± 1455.4	52.67643 ± 997.8872	-1447358 ± 45737
Highest average	-732187.2 ± 7738.5	493200 ± 8001.25	-131.8668 ± 16724.34

The regression model produced are worse than the previous section even when the regression model describes the same set it was generated for. This is seen in R-square and R-adjusted values. The mean is far off from the value zero and the 95% confidant interval is quite wide which entails lack of accuracy.

None of those three models have a high accuracy of predicting the given sets. This is telling us that the value of predictions has different effect on different sets.

4. Development of the additional segmentation model is explained and the model is shown.

In this section, I sorted the whole sets of house prices from the highest house price to the least house price. Them I divided the sorted list into the following 3 sub-sets:

- The highest 10% of house prices
- The median 80% of house prices
- The lowest 10% of house prices

The Highest 10% of House prices

After going through the elimination process the predictions used are:

$$lm(formula\ =\ price \sim bedrooms + bathrooms + sqft\_living + floors + view + grade + sqft\_above + yr\_built + zipcode + lat + long)$$

With

$$R^2: 0.552\ ||\ Adjusted - R^2: 0.5489$$

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.169e+07  2.223e+07   1.876  0.06085 .  
bedrooms    -6.586e+04  1.321e+04  -4.987 6.78e-07 ***
bathrooms    8.956e+04  1.845e+04   4.855 1.32e-06 ***
sqft_living  2.355e+02  2.052e+01  11.479  < 2e-16 ***
floors      -6.921e+04  2.591e+04  -2.671  0.00763 ** 
view         8.005e+04  7.842e+03  10.208  < 2e-16 ***
grade        7.477e+04  1.266e+04   5.908 4.23e-09 ***
sqft_above   1.034e+02  2.182e+01   4.740 2.33e-06 ***
yr_built    -3.809e+03  4.155e+02  -9.168  < 2e-16 ***
zipcode     -3.510e+03  2.579e+02 -13.612  < 2e-16 ***
lat          9.157e+05  1.666e+05   5.497 4.49e-08 ***
long        -2.176e+06  1.275e+05 -17.076  < 2e-16 ***
```

The Median 80% of House prices

After going through the elimination process the predictions used are:

$$lm(formula\ =\ price \sim bedrooms + bathrooms + sqft\_living + floors + view + grade + sqft\_above + yr\_built + zipcode + lat + long)$$

With

$$R^2: 0.6047||\ Adjusted - R^2: 0.6043$$

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -9.633e+06  2.004e+06  -4.807 1.55e-06 ***
bedrooms      -7.822e+03  1.254e+03  -6.236 4.63e-10 ***
bathrooms      1.936e+04  2.176e+03   8.898  < 2e-16 ***
sqft_living    5.940e+01  2.452e+00  24.221  < 2e-16 ***
sqft_lot       1.878e-01  2.592e-02   7.246 4.53e-13 ***
floors         3.128e+04  2.113e+03  14.805  < 2e-16 ***
view           2.503e+04  1.511e+03  16.569  < 2e-16 ***
condition      1.919e+04  1.582e+03  12.133  < 2e-16 ***
grade          6.305e+04  1.479e+03  42.626  < 2e-16 ***
yr_built      -1.776e+03  4.598e+01 -38.636  < 2e-16 ***
zipcode       -1.048e+02  2.051e+01  -5.110 3.27e-07 ***
lat            4.853e+05  7.098e+03  68.362  < 2e-16 ***
sqft_living15  4.479e+01  2.449e+00  18.286  < 2e-16 ***
```

The Lowest 10% of House prices

After going through the elimination process the predictions used are:

$$lm(formula = price \sim bathrooms + sqft\_living + view + condition + grade + lat)$$

With

$$R^2: 0.2236 || Adjusted - R^2: 0.2207$$

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.337e+06  3.549e+05  -6.585 6.14e-11 ***
bathrooms    9.518e+03  1.962e+03   4.851 1.35e-06 ***
sqft_living  2.171e+01  2.835e+00   7.659 3.23e-14 ***
view         1.350e+04  5.141e+03   2.625  0.00874 ** 
condition    9.601e+03  1.142e+03   8.405  < 2e-16 ***
grade        9.983e+03  1.230e+03   8.115 9.59e-16 ***
lat          5.065e+04  7.428e+03   6.819 1.29e-11 ***
```

Compare each set with the other two and determine how good the model describing each set. 95% confidence interval were used to obtain the range from the mean. For each set it was halved into two subsets for train set and for test set. The following table is meant to compare how the three regression models holds when used to describe different sets:

Price Range	**Lowest 10%**	**Median 80%**	**highest 10%**
Lowest 10%	-6.111695 ± 211.9601	-236620.2 ± 355	-1026239 ± 4009.5
Median 80%	66297.05 ± 510.865	5.797595 ± 250.1572	-535881.2 ± 3582.15
highest 10%	-732187.2 ± 7738.5	50841.91 ± 735.82	143.4041 ± 2844.053

This is the best regression models yet that describe data points quite good. This is shown when conducting a t-test when compare between the actual data and the train ones. It gives values closer to zero as well as the range of 95& confidant interval is relatively tight. This is not only in describing the set it was trained on, but also its good describing other two sets.

5. Explanation of prediction results, including RMSE values for each model.

By using RMSE, we can indicate how good a regression model with high confidant. The less RMSE value, the better the regression model is. The ideal situation would be RMSE equal to zero which indicates a perfect fit. The larger it is the less relevant to the actual data points. This is what kaggle is using to preduce a score value. RMSE is represented by the following equation:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Kaggle websites generated the following RMSE values for each model:

1. Full set of predictions for house prices: **181158.48**
2. After eliminated set of predictions for house prices: **184116.65**
3. Zip code with the lowest average price: **343258.66**
4. Zip code with the median average price: **493944.29**
5. Zip code with the highest average price: **1270828.38**
6. The lowest 10% of the house prices: **403171.65**
7. The median 80% of the house prices: **207176.96**
8. The highest 10% of the house prices: **318924.43**

My best RMSE value that is reported by Kaggle was when I used the full set of predictions for house prices and it gives:

$$\boldsymbol{RMSE = 181158.48}$$

5. Data cleaning is described and performed adequately.

Going through the prosses of generating a regression model, I made sure to follow those steps:

1. eliminate any predictions than has no effect on the model.
2. eliminate any predictions that has little to contribute to the model because adding such a prediction will be make the model over fitting the data points.
3. Each time the train/test set is conducted 100 times to gather more data points and therefore have a better data description when t-test is implemented on it.
4. Making sure any reduced model is not only tested on the set it was trained on, but also other sets from other categories to see how will the model holds.