Statistical Data Mining I - Homework 1

Tuesday, September 17, 2019 9:19 PM

Nikita Goswami
UBIT Name : Nikitago
E-mail : nikitago@buffalo.edu

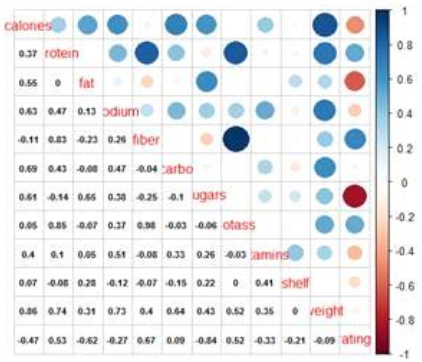
Q1) CEREAL DATASET: Exploratory Data Analysis and Pre-processing of Data

- There are 16 attributes and 77 different types of cereals in the data
- Name, Manufacturer and Type columns are factor type (qualitative). All the other variables are integer and numeric type(quantitative)
- Nutrition values in the dataset is being defined per serving but manufacturers are defining serving with different weight and cups. For comparing the nutrition values, I have normalized all the nutritional ratings per cup.
- There are some missing values in the dataset denoted as '-1'. Substituting it with NA and would be ignoring these values while creating plots and the model
- There are outliers present in nutrients like sodium, potassium, calories and vitamins but I would not be treating them as they look legitimate values and I would not want to bias the data by removing them from the analysis. Also since we have very less data, removing a single data will be very costly

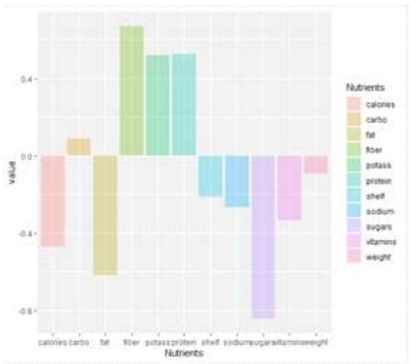
Observations :

- Prominent Manufacturers are 'Kellogs' and 'General Mills', in terms of the number of cereals manufactured by them
- 'Cold' type of cereals are significantly greater in number than the 'Hot' type.
- 'Shelf' attribute has 3 values based on the position where cereal is kept. It is a numerical value in the dataset but taking it as numerical data may bias the model towards cereals on shelf 3. Thus taking it as a qualitative variable for the analysis

Correlation Matrix

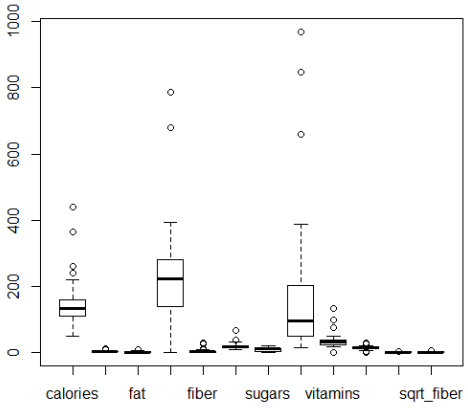


Correlation matrix in a bar graph form



a. Cereals with good fiber, potassium and protein have higher ratings
b. Cereals with higher level of Sugar, fat and calories have lower ratings(Though with calories and carbo, we do not have lot of data points in the higher side for these variables to establish the exact relationship)
c. Name of cereal, shelf , manufacturer and type are categorical variables. We can study them to see if they can add value to our model.

Outlier Treatment

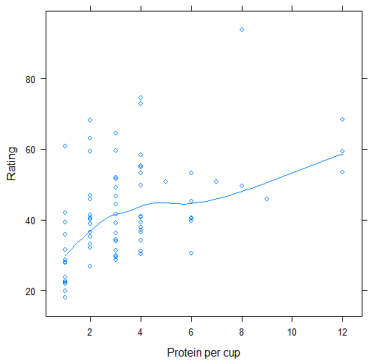


The Box Plot shows outlier in many nutrient values . Since we have very less data points to do our analysis and on further study the outlier values look like legitimate values.
So removing them or changing the values can affect our model negatively.

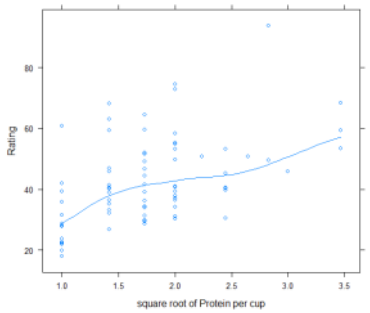
Studying Independent attributes

Sugar has a negative relationship with Rating. After sugar reaches it's threshold value (around 14), the rating is constant. The relationship thus is not linearly negative but applying transformations to make the relationship linearly negative is giving unwanted attention to few points which have less sugar and high rating. Thus I would not go about transforming sugar values for my model.

Study of Protein v/s Rating

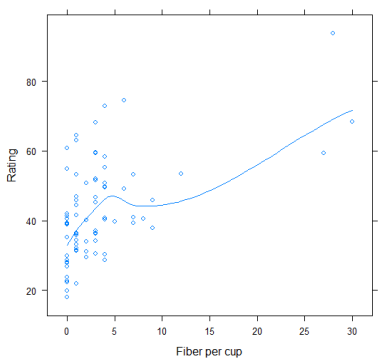


Protein has a positive relationship with rating but when protein>8, few points in the high range are having a lot of say in the linear relationship. To even out the data, I would apply the square root transformation to protein and use this value in my model.

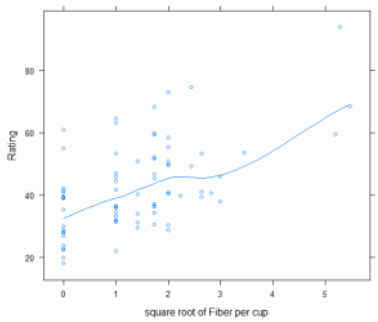


Study of Fiber v/s Rating

First.



The plot is linear when **fiber**>15 but few points in that range are having a lot of say in the linear relationship. When fiber is between 5 and 10 range, rating is going down which is an unexpected trend. We cannot apply a log transformation to Fiber as the min value is 0. Applying, square root transformation is giving a good line.



Plot between **Fat** and Rating shows that the cereals with no fat have a higher rating and as fat value increases, the rating is static.
Comparing the adjusted R-squared estimate on including sodium and square root of sodium, The square root value is giving an higher estimate

```
Call:
lm(formula = rating ~ sqrt_fiber + sqrt_protein + sqrt_sodium +
    sugars + fat + calories + carbo + potass + vitamins, data = cereal_data2,
    na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7265 -1.6325 -0.2199  2.2027  6.9438

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.09123    2.44463  18.445  < 2e-16 ***
sqrt_fiber    4.21759    0.89707   4.702 1.42e-05 ***
sqrt_protein  6.76546    1.67028   4.050 0.000141 ***
sqrt_sodium  -0.73969    0.09078  -8.148 1.77e-11 ***
sugars       -1.14712    0.22024  -5.208 2.16e-06 ***
fat          -1.71453    0.54914  -3.122 0.002694 **
calories     -0.02980    0.04828  -0.617 0.539198
carbo         0.27432    0.25428   1.079 0.284722
potass        0.02172    0.00645   3.368 0.001287 **
vitamins     -0.05398    0.01981  -2.725 0.008272 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.701 on 64 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.939,     Adjusted R-squared:  0.9305
F-statistic: 109.5 on 9 and 64 DF,  p-value: < 2.2e-16
```

Q2. a)Which predictors appear to have a significant relationship to the response.

1. Cereals with good fiber, potassium and protein have higher ratings
2. Cereals with higher level of Sugar, fat and calories have lower ratings(Though with calories and carbo, we do not have lot of data points in the higher side for these variables to establish the exact relationship)

3. **What does the coefficient variable for sugar suggest?**
   For every unit increase in sugar, the rating is decreasing by -1.14

4. **Use the “:” and “*” symbols to fit model with interactions. Are there any interactions that are significant?**

- ":" takes into account interation by multiplying the two attributes. "*" takes into account addition of separate attributes and then multiplication as well.
- Fiber and Protein have a positive relationship with Rating and look related. Adding interaction Sugar:Fat
- Sugar and Fat have a negative relationship with Rating and are related. Adding interaction Fiber:Protein
- Since we are already including all the attributes separately. I am using interaction variables with ":" to fit the model
- R-squared statistics estimates that 95.89% of the changes in the response can be explained by these particular set of attributes.
- Adding interaction between other variables is increasing the R square value further but I believe that it may lead to overfitting and so I would not include it in my model. If test data was available to test the model with various attributes , the decision making would be more clear.

```
Call:
lm(formula = rating ~ sugars:fat + sqrt_fiber + sqrt_protein +
    sqrt_sodium + sugars + fat + calories + carbo + potass +
    vitamins + fiber:protein, data = cereal_data2, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6138 -2.1831  0.2353  1.8534  6.2523

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   45.346852   2.191682  20.690  < 2e-16 ***
sqrt_fiber     5.271404   0.840687   6.270 3.87e-08 ***
sqrt_protein   7.750884   1.487548   5.211 2.28e-06 ***
sqrt_sodium   -0.739220   0.080602  -9.171 3.78e-13 ***
sugars        -0.915926   0.227843  -4.020 0.000160 ***
fat           -3.542143   0.902160  -3.926 0.000219 ***
calories      -0.108297   0.047926  -2.260 0.027366 *
carbo          0.655127   0.252204   2.598 0.011712 *
potass        -0.007255   0.009936  -0.730 0.468045
vitamins      -0.056205   0.017637  -3.187 0.002255 **
sugars:fat     0.229424   0.065324   3.512 0.000836 ***
fiber:protein  0.058221   0.019209   3.031 0.003555 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.26 on 62 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.9542,    Adjusted R-squared:  0.946
F-statistic: 117.3 on 11 and 62 DF,  p-value: < 2.2e-16
```

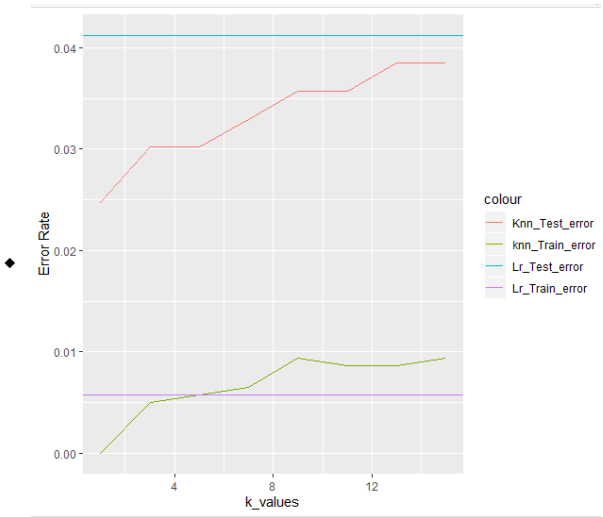
Adjusted R-squared statistics estimates that 94.6% of the changes in the response can be explained by all set of predictors in data.

3. **Are there any interactions that are significant?**
   Interactions between Sugar:Fat and Fiber:Protein look promising and increase the R-squared estimate

Question 3) Zip Code Data : Comparing Linear Regression and K-nearest neighbors for classification

Observations :

- Train data has 7291 rows and 257 columns. Test Data contains 2007 rows and 257 columns
- Converting the data from matrix type to Data Frame type as the lm() function for fitting linear model runs on Data Frame objects.

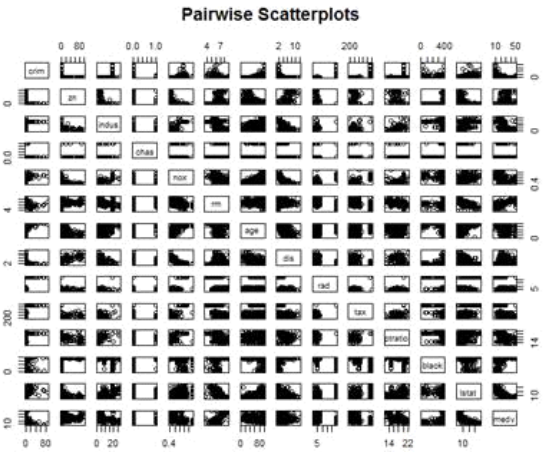


- The results show the best accuracy when value of K is 1. One of the reasons why the closest neighbor is successful in giving the best results may be due to high dimensions (256 attributes may be good enough to distinguish ‘2’ from ‘3’). There is a risk of overfitting here as we have fewer data points. (it is 1389!).
- **KNN is giving better performance in classifying data than linear regression.** The reason is that linear regression is suitable for predicting continuous values. We have taken an assumption that any value greater than 2.5 will be treated as 3.

Question 4) Study of Boston Housing Dataset from MASS Library

- Observations are made in 506 neighborhood of Boston for 14 predictor attributes

1. **Make a matrix of pairwise scatter plots**



The scatterplot gives us lot of insights into the Boston suburbs data. Some of them are :

1. Only the places where there is no residential land (zn is close to zero), the crime rate is high. In the residential areas, the crime rate is low.
2. In the suburbs where the houses are comparatively new, the crime rate is less. In suburbs where the house are extremely old, the crime rate is varying and is more than triple in few suburbs.
3. In the suburbs which are closer to the employment areas(dis), the crime rate is higher compared to areas which are away.

b) Are any of the predictors associated with the per capita crime rate.

Correlation matrix of crime rate v/s all other attributes

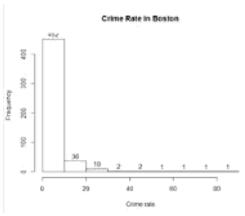
- High level of rad - index of accessibility to radial highways contain the highest level of cri - per capita crime rate in town.
- High level of tax - full-value property-tax rate per $10,000 has a correlation with the per capita crime rate in town
- lsat, nox and ndus are other attributes having some amount of linear relationship with per capita crime rate.
- Medv and Black have the max inverse relationship with per capita crime rate

c) Do any of the suburbs appear to have particularly high Crime Rates

There are some neighborhoods where the crime rate is very high

2% of suburbs in Boston have crime rate above 25

0.8% of suburbs in Boston have crime rate above 50



Tax Rates

There are many suburbs with high tax rate in Boston

72% of suburbs in Boston have tax rate below 500

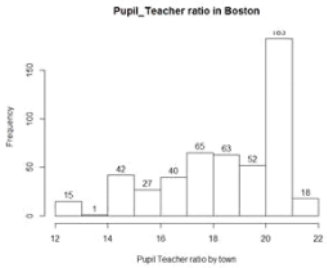
27% of suburbs in Boston have tax rate above 500

a. **Analysis of Pupil-Teacher ratios in the suburbs of Boston**

There are many suburbs with higher pupil teacher ratio compared to other suburbs

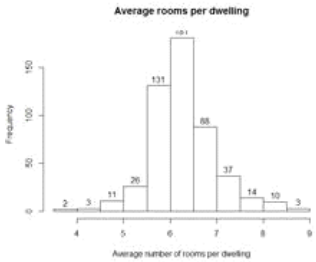
11% of suburbs in Boston have pupil teacher ratio below 15

88% of suburbs in Boston have pupil teacher ratio above 15



b. **A)How many suburbs average more than 7 rooms per dwelling**

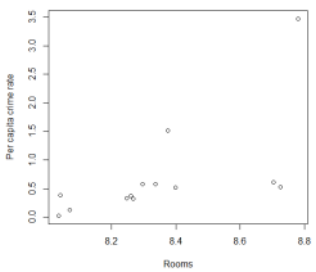
64 suburbs have more than 7 rooms per dwelling



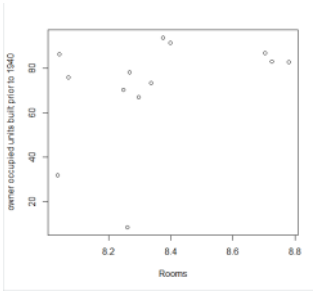
B) How many suburbs have more than 8 rooms per dwelling

13 suburbs have more than 8 rooms per dwelling

My hypothesis is that as the number of rooms are increasing, the suburbs would mostly have educated and wealthy people. Looking at the crime rates in these areas, we see that crime rate varies with an exceptionally high rate in one of the suburbs.



Most of the houses in these suburbs are old.



We would assume that more number of rooms would mean higher price of the houses. The plot shows the same, the value of homesis having a positive relationship with rooms.