STA363Proj3 - Formal Report

Rickey huang

4/19/2021

Abstract/ Executive Summary

Section 1: Data Cleaning

In this section, the correlations between data and the meanings of the data themselves are explored. Some dulicaptive variables for the analysis is deleted.

Section 1.1: Refining the Variables

By computing the correlation matrix of the numeric varaible in the data, the correlation between every pair of numeric variables in this data is explored. There are some highly correlated variables. The correlation between the variables *"Calories.from.Fat"* and *"Total.Fat"* is 0.99957642, and the possible reason for this strong correlation is the fact that one gram of fat normally containes 9 calories, which is confirmed by the data. Hence, the *"Calories.from.Fat"* is removed from he data set. However, since the variable *"Calories"* is the number of calories in per serving, it should be kept in the dataset because it measures the total calories in the serving but not only the calories from the fats of the serving, which means it is different from the variable *"Calories.from.Fat"*. The correlation between the variables *"Total.Fat"* and *"Total.Fat....Daily.Value."* is 0.99970351 for the reason that the *"Total.Fat....Daily.Value."* is measure by divide the *"Total.Fat"* by the total recommended of fat and then times 100. For this reason, the variable *"Total.Fat....Daily.Value."* can be deleted from the data set. For the similar reason, the daily values *"Saturated.Fat....Daily.Value."*, *"Cholesteril....Daily.Value."*, *"Sodium....Daily.Value"*, *"Carbohydrate....Daily.Value."*, and *"Dietary.Fiber....Daily.Value."* should be removed since they have correlations of 0.9992613, 0.99985282, 0.999919583, 0.99961372, and 0.98592990 with their corresponding actual values.

Finally, since the *"Item"* variable is a unique variable that represents the name of each serving at McDonalds, it could be used as an identifier for the each observation, but it is not useful for exploring the relationship with *"Calories"*, so that it would be removed from the dataset in order to do the further exploration.

After cleaning and refining the data set, we get a data with 259 observations and 17 vairables. Among those variables, two of them are categorical variables *"Category"* and *"Item"*, while the rest of variables are numeric variables that measures the size and contents in the servings from McDonalds.

Section 2: Modeling Calories

Section 2.1: Distribution of the Response Variable

In order to show the trend in the response variable *"Calories"*, a historgram for it is created as shown in the Figure 1. From the histogram, the mean 362.4324 is indicated by the dashed light blue vertical line in the visualization. By the relative relation between the mean and the median reveal from the histogram and the density curve in the histogram, the distribution of the Calories should be identified as skewed to the right.

Section 2.2: the Tree1 Model

Section 2.2.1: Reasons for Choosing the Regression Tree model

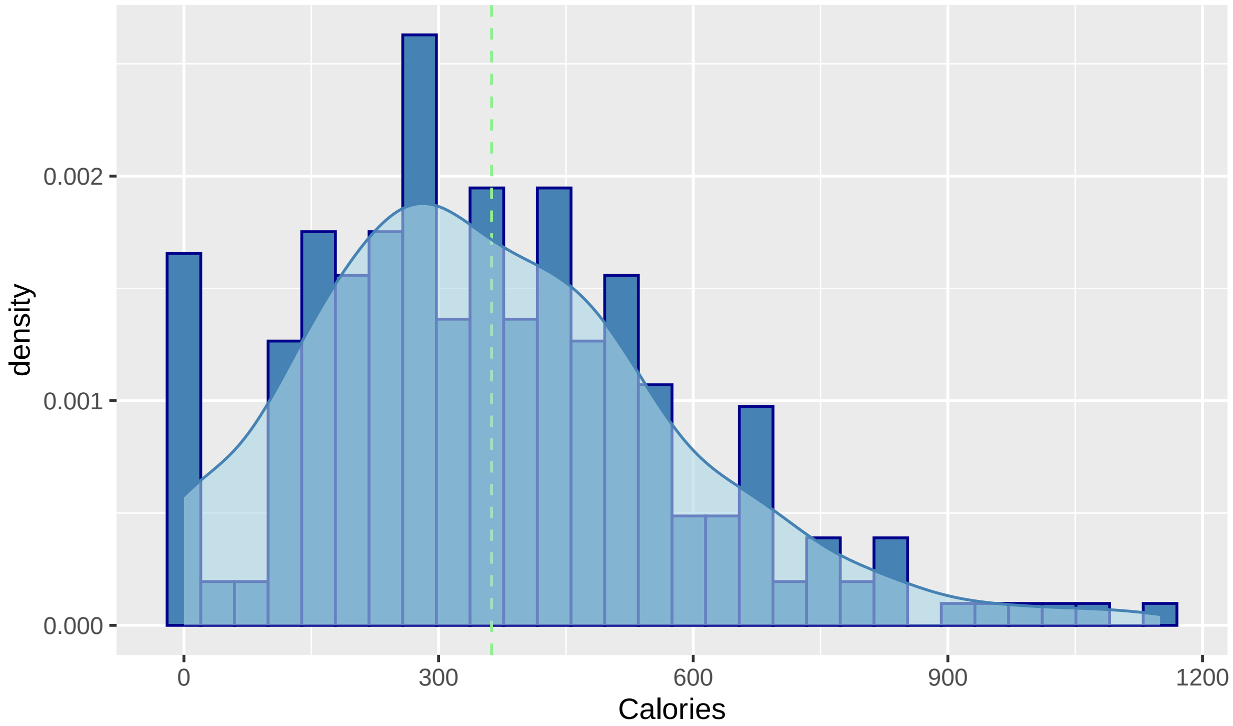

Figure 1: Histogram for Calories

Since the client wants to understand the relationship between different features and the amount of Calories in the food and also wants a clear explanation, a regression tree model would be a good choice for our client to use to create a great visualization to explore and show the relationship. Since our response variable *"Calories"* is a numerical variable, we could use the regresion tree to do show the numerical relationship between the response variable and features.

Section 2.2.2: Training the Tree Model Using All Features (fullTree1)

First, all features are included to fit the tree model fullTree1. The fullTree1 model has 8 splits and the Root Node Error (RNE) for the fullTree1 is 48872. In order to evaluate how this tree model performs by creating each split, the cp table for the fullTree1 model is computed as shown in the Table 1. From the cp table, the percentage change from the RNE (xerror in the table) by creating each split is shown row by row. When the 8^{th} split is created, the percentage change from the RNE for the test MSE is 16.34231%, so that according to the Forumula 1, the test MSE for the fullTree1 is $48872 \times 16.34231\% = 7986.814$.

$$testMSE = RNE \times xerror \tag{1}$$

Table 1: The cp table for the fullTree1

CP	nsplit	rel error	xerror	xstd
0.5988049	0	1.0000000	1.0039544	0.1013384
0.1312561	1	0.4011951	0.4271343	0.0514122
0.0761375	2	0.2699390	0.3635852	0.0454102
0.0444821	3	0.1938015	0.2379931	0.0256588
0.0215682	4	0.1493194	0.2196494	0.0212518
0.0192588	5	0.1277512	0.1965999	0.0206247
0.0116562	6	0.1084923	0.1800554	0.0205555

CP	nsplit	rel error	xerror	xstd
0.0111753	7	0.0968361	0.1650205	0.0197030
0.0100000	8	0.0856608	0.1634231	0.0196925

Section 2.2.3: Pruning the fullTree1 to create the Tree1 Model

each split in the fullTree1 model is created by minimizing the RSS of the model. However, the fullTree1 may overfit on some points which would results an inaccurate clarification for the relationship. In order to avoid this overfitting problem that may exist in the fullTree1 model, the cost complexity pruning technique is used to prune the fullTree model to produce a more appropriate tree model. In stead of minimizing the RSS, the cost complexity metric (C_α) is minimized. The formula for C_α is expressed in the Formula 2, where $\alpha\,|T|$ is the penalty term add to the RSS to do the pruning, α is the tuning parameter, and T is the number of leaves in the tree model.

$$C_\alpha = RSS + \alpha\,|T| \tag{2}$$

Again, the cp table is used to choose a convincing tuning parameter by looking one with a smallest test MSE. Also, for the convenience for explanation, we want a model with as least splits as possible. Using this rule, and from the Table 1, a 7-split tree model should be chosen since one more split after the 7-th split only result in a 0.002 reduction in xerror, which would not bring us much new information about the relationship with *"Calories."*. The percentage change from the RNE for the 7-th split is 0.1650205, which would produce a test MSE of 48872 × 16.50205% = 8064.882. Then the tuning parameter chosen is 0.0111753. Also, the cp plot in the Figure 2 that visualizes the relationship between the percent change from RNE and cp values also confirms this choice of α.

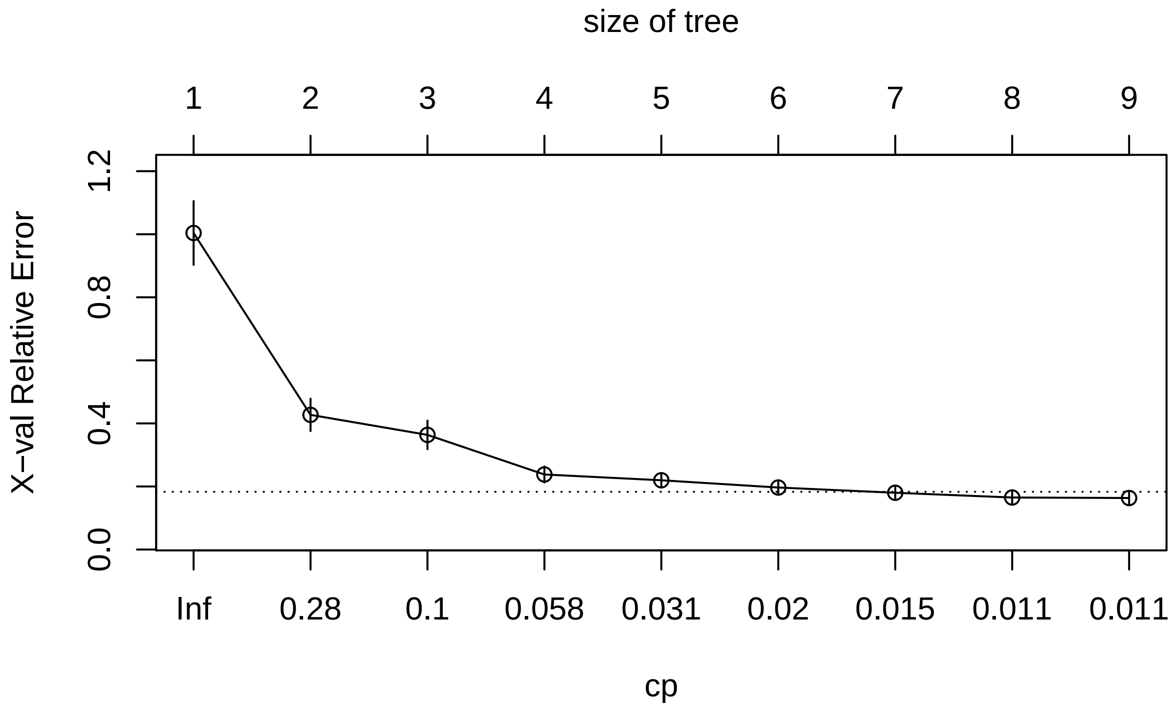

Figure 2: cp plot for fullTree1

With $\alpha = 0.0111753$, the Tree1 model could be fitted. The visualization of the regression tree model Tree1 is shown in the Figure 3.

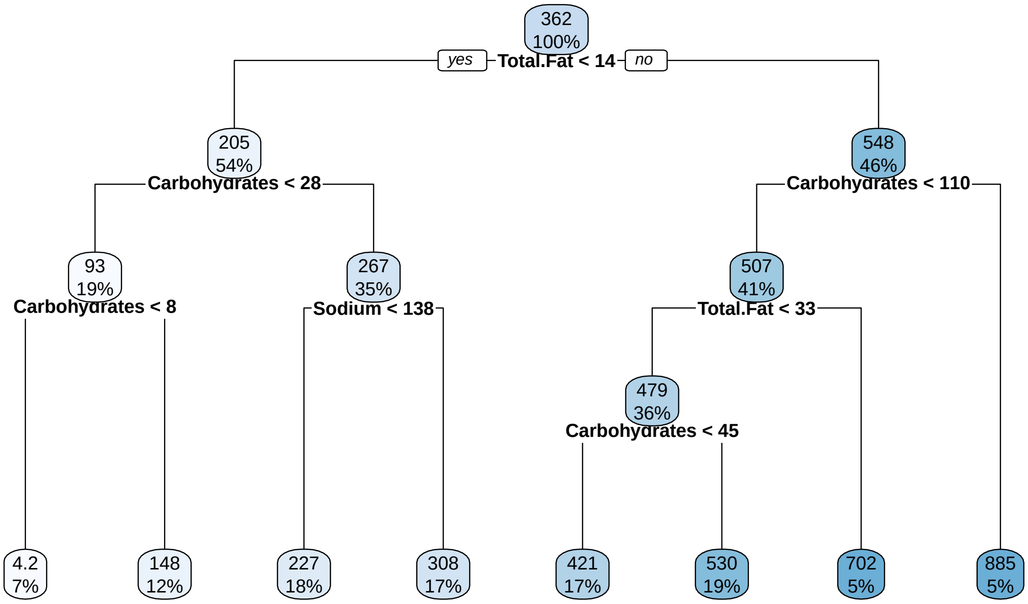

Figure 3: Visualization for Tree1

Section 2.2.4: Infromation revealed by the Tree1 model

From the visualization in the Figure 3, each split is decided by minimizing the cost complexity metric C_α. For example, the reason that the *"Total.Fat"* is used as the first split and it splits at 14 is that comparing to using other features to make the first split or spliting at other values, the first split on *"Total.Fat"* at 14 would result an optimal cost complexity metric. Similarly, the further splits are made in this way as well.

The visualization also uses the colors to indicate the amount of calories of the servings. The higher calories the serving are, the color for that leaf would be darker. Hence, a high value of *"Total.Fat"* and *"Carbonhydrates"* would result in a relatively high calorie serving, while a low value of *"Total.Fat"*, *"Carbonhydrates"*, and *"Sodium"* is more likely to be related with low calorie foods. The detailed quantified splits using combinations of these four features can be explored in the visualization. For example, if a serving has *"Total.Fat"* content of 15 and *"Carbonhydrates"* content of 93, the prediction for the calories in this food would be 530, which is a relatively high colaries.

In order to evaluate the accuracy of this tree model, the training and test RMSE are calculated. The training RMSE is the square root of the RNE of the model, so the training RMSE of the Tree1 model is $\sqrt{48872} = 221.0701$, while the test RMSE could be calculated by taking the square root of the test MSE that we obtained before using the Formula 1. Hence the test RMSE is $\sqrt{8064.882} = 89.80469$.

Section 3: Modeling Category

Section 3.1: Distribution of the Response Varaible

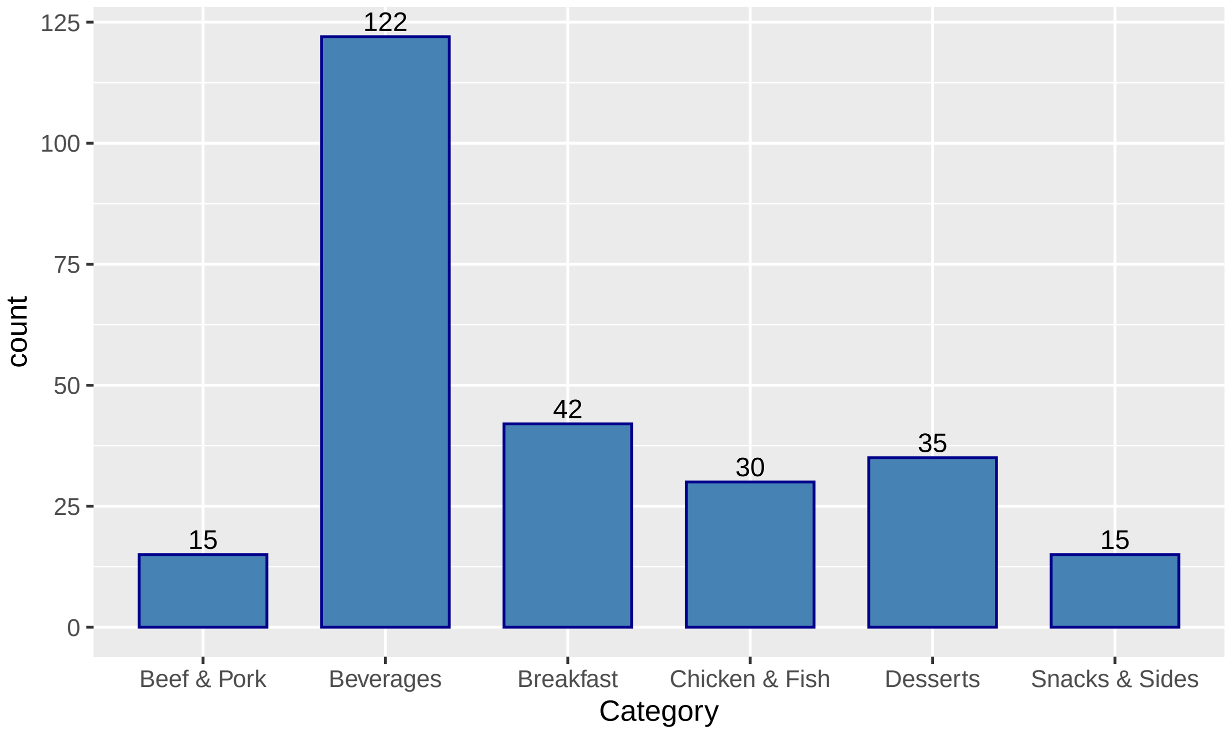

Figure 4: the Distribution of the Responce Variable Category