ALY 6040 Data mining  
Module 3: Icecubed

Professor: Justin Grosz

Chongfan Bao CPS in Informatics

Northeastern University

06/13/2021

**Abstract**

For this project, I am going to implement feature engineering, clustering and principal component analysis (PCA) to icecubed data set. This is as so to develop data model for machine learning or logistic regression and reduce the running time of model. In addition, by exploration data analysis, I am able to understand which several variables are similar or important for our data. Base on this, I can reduce or increase the dimension of data set and divide the data into different categories.

*Keywords:* Feature engineering, PCA, Clustering

**Introduction**

# Desserts are so admirable for many people, and everyone basically likes to eat desserts after meals or at ordinary times. This is not only because the dessert is delicious, but also because it can make us happy. Especially in hot weather, ice cream is always a favorite of people. In this assignment, in order to understand people’s preference for different ice creams and the number of purchases, I will explore the data set by classify or cluster data to contribute to machine learning modeling to the best of my ability.

# Data overview

The icecubed consists of 10000 observation and 12 variables, which include 5 numerical variables and 7 characteristic variables. From the figure 1, I am able to notice the number of classifications of each variable and description statistics of the numerical variable. Meanwhile, I also find that there is not any missing data. Luckily, I don’t need to deal with the missing values. However, due to the numeric variable requirements of the machine learning models such as tree decision and logistic regression, I need to change the characteristic variable into binary variables.

# Data preprocessing and modeling

Before data modeling, I need to implement exploration data analysis and data clean so as to increase the efficiency of model and ensure the accuracy of results. First of all, I am going to look for if there is any missing data in each column. This is in order to ensure effective and smooth data analysis and modeling, as well as the accuracy of data results. If there is any missing in the data set, some function in the Rstudio cannot deal with the missing data. Next, I will check the duplicate data. Many would probably not immediately realize the importance of finding and removing duplicate data, duplicate data can create chaos, time wasting and the accuracy of outcomes and model that might, eventually, cost your business a considerable amount of money. Fortunately, in this data set, I don’t find any missing data and duplicate data (The results display in Appendix figure 2). I don’t need to carry out data cleaning. This is also because the data have been cleaned. Now, it is time to conduct one-hot encode and logistic regression modeling. Because the logistic regression function and correlation analysis requires the variable to be a numeric variable, the first thing that I need to do is to change the raw characteristic data into numeric variable and build a new data frame that there is only numeric variable. One-hot encode is important way that people transform the data into the binary numeric format, which enhances the computational power and the efficiency of the machine learning algorithms. Another feature engineering is Binning. In this model, I use “cut” function to divide the variable named “How many desserts do you eat a week” into three groups: low, mid and high based on 1st quarter, median and 3rd quarter. The results in figure 5 show 3200 low, 5336 mid and 1464 high. This is in order to precisely research different target customer’s consumption habits. In addition, binning allows the model to contain missing data and other special calculations, to control or reduce the influence of outliers on the model, to solve the problem of different scales between features, and to make the weights of coefficients in the final model comparable.

Logistic regression is an appropriate regression analysis to be performed when the dependent variable is binary. Logistic regression is a predictive analysis that is used to describe data and explain the relationship between a dependent binary variable and one or more independent variables of nominal, ordinal, interval or ratio levels. In this module, I will select some variable to implement the logistic regression. This is in order to find which variables exist relationship with dessert purchase.

Clustering and principal component analysis (PCA) both belong to dimension reduction. They are in order to find the variables that are similar or work out which variables make one group different from another and aggregate them into different categories. Furthermore, dimension reduction can reduce the difficulty of processing high-dimensional data, reduce the complexity of subsequent calculations, remove noise and redundant data, and reduce the loss of information at the same time. Low-dimensional data is easier to understand and visualize than high-dimensional data. However, both methods work best with numerical data. Therefore, I have to select the numerical variables from raw data to build a new data frame. This new data frame is made of five numeric variables and 10000 records.

**Data Analysis**

One-hot encode is used to quantify features. In figure 3, I transfer all character numerical into binary. This is in order to fit logistic regression and plot correlation matrix to research the correlation of each variable. For example. The sex column in raw data set is labeled as “male” and “female”. In the new data frame I created will leverage the “10” as “male”, “01” as “female”. Eventually, the value will vary from character into numerical and the numerical variable will keep the original form. At this time, I can develop a logistic regression and plot correlation matrix. According to figure 4, “Ice.Cream.Products.Consumed.Per.Week” and “How.many.desserts.do.you.eat.a.week”, “gendermale” and “Household.IncomeNot.Reported”, “Donate.Date2019.07.07” and “Deposit.Amount” exit strong correlation. Based on this, I will conduct logistic regression for these variables. From figure 6, we can know each variables intercepts and P- value, which is helpful to verify our hypothesis. If the p-value for a variable is less than your significance level, your sample data provide enough evidence to reject the null hypothesis for the entire population. In K -means cluster analysis, the number of groups pre-specified by the analyst. It classifies objects in multiple groups so that the objects in the same cluster are as similar as possible, and the objects in different clusters are as dissimilar as possible. However, I have been specified to divide the data into three groups. By the calculating and compare with plots in figure 7 and 8, I find that two clusters are the best. If there are three clusters, I notice that the cluster 1 almost includes the cluster2.

As for PCA, in figure 9, I can notice that the cumulative contribution rate of the first three principal components has reached 83.318%. The other one principal components can be discarded to achieve the purpose of dimensionality reduction. In addition, based on percentage of variables, Dim.1 explains 35.418% of the total variance, which means that nearly one-thirds of the information in the dataset can be encapsulated by just that one Principal Component. Dim.1 explains 29.587% of the variance. Dim.3 explains 18.314% of the variance. So, by knowing the position of a sample in relation to just Dim.1, Dim.2, and Dim.3, I can get a very accurate view on where it stands in relation to other samples, as just Dim.1, Dim.2, and Dim.3 can explain 83.818% of the variance.

**Conclusion**

All in all, each model has own pros and cons. For one-hot encode and Binning, it can help me to refine variables and create new variables to research and optimize models. It also allows us to reformulate non-linear problems as linear problems, which is of great help to us in deep data mining. Clustering and PCA analysis play a role in dimensionality reduction, shrink huge data sets, and organize observed data into meaningful structures, which help speed up machine learning algorithms。

**References**

Deepika Singh, (2019). Encoding Data with R. Retrieved from https://www.pluralsight.com/guides/encoding-data-with-r

Dimension reduction | Principal component analysis. Retrieved from https://www.cnblogs.com/leezx/p/6120302.html

James, L., (2018). Logistic Regression in R Tutorial. Retrieved from https://www.datacamp.com/community/tutorials/logistic-regression-R

Kai (2018). Principal Components Analysis. Retrieved from https://www.bioinfo-scrounger.com/archives/608/

UC Business Analytics R Programming Guide. K-means Cluster Analysis. Retrieved from https://uc-r.github.io/kmeans\_clusteringMarco P., (2011). Ames Iowa Housing Data. Retrieved from https://www.kaggle.com/ marcopale/housing

Appendix A

**Figure1:** *Data summary*

Table

Description automatically generated

**Figure2:** *Check missing and duplicate data*

*Graphical user interface, text, application, email

Description automatically generated*

**Figure3:** *One hot encode*

*Table

Description automatically generated*

**Figure4:** *Correlation matrix*

*Chart, scatter chart

Description automatically generated*

**Figure5:** *Binning*

*Graphical user interface, text

Description automatically generated*

**Figure6:** *Logistic regression*

*Table

Description automatically generated*

**Figure7:** *calculate clusters*

**Chart, line chart

Description automatically generated**

**Figure8:** *plot cluster*

*Chart

Description automatically generatedChart

Description automatically generated*

**Figure9:** *Eigenvalues PCA*

*A picture containing table

Description automatically generated*

**Figure10:** *Variable PCA*

*Chart

Description automatically generated*

Appendix B

R code

Text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generatedText

Description automatically generated