Data Memo

Xilong Li

2022-04-10

- An overview of your dataset

What does it include?

It is a sales report that contains: 1) the commodities sales data, including the price, brand, sold time, categories, etc
2) the corresponding data from the customer who bought this good, including their age, approximate income, gender, etc

Where and how will you be obtaining it? Include the link and source.

This data set was obtained while I was working as a data analyst in a local grocery company in my hometown. The company permitted the use of this data since it was engaging a academic research with the Tsinghua University, and thus I was able to download the data and store it on my computer.

About how many observations? How many predictors?

There are approximately 6000 observations, and 15 predictors.
However, I would have to reduce its complexity in order to make the establishment of model easier.

What types of variables will you be working with?

The variable will primarily be categorical variable, discrete variable, and continuous variable, which will include predictors such as gender, age, income, sales price, etc.

Is there any missing data? About how much? Do you have an idea for how to handle it?

There are many missing data about the customer's name, age, gender , and income. But I would primarily not use those predictors anyway.
Also, the variables such as age and gender do not have too many missing data, so I can simply exclude those missing data.

- An overview of your research question(s)

What variable(s) are you interested in predicting? What question(s) are you interested in answering?

1) The question I am interested in: is to predict how the age and gender of the customer will affect their buying frequency, total buying amount, or types of commodity.

2) It seems to me that older people might have higher buying frequency and amount, since they are usually the ones who take care of their grandchildren in China.

Name your response/outcome variable(s) and briefly describe it/them.

possible response variables: 1) sales amount — total amount that is recorded each time the customer enter the grocery store. 2) buying frequency — how many times a week/month does this person goes to the store 3) commodity category — category of this commodity

Will these questions be best answered with a classification or regression approach?

The answer might be best answered by regression approach, since the primary predictors might include age, income, etc.

Which predictors do you think will be especially useful?

Age and gender might be especially useful.

Is the goal of your model descriptive, predictive, inferential, or a combination? Explain.

For now, I am deciding to make the goal of my model predictive, in order to show what kind of commodity might be popular among certain age of customers, so that the company can use the prediction to decide the flow in & out of certain commodities.

- Your proposed project timeline

When do you plan on having your data set loaded, beginning your exploratory data analysis, etc?

I plan to load and start in this week. Even though I already had the data, I still need to exclude many unrelated variables.

Provide a general timeline for the rest of the quarter.

I plan to first load the data and try to make some regression. But I am honestly not sure right now because I haven't learned anything from classes, and I don't know what exactly I can do with the data.

- Any questions or concerns

Are there any problems or difficult aspects of the project you anticipate?

Perhaps it would be difficult to establish model including so many variables.
It is also possible that some variables might have too many missing values that I have to stop using them.

Any specific questions you have for me/the instructional team?

I think it would be helpful to do a overview about what exactly are we gonna learn this quarter in a organized way so that we can know what we can do with our projects.