Berkeley Engineering | BerkeleyHaas

PROFESSIONAL CERTIFICATE IN MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Office Hour #5 with Matilde D'Amelio

April 7, 2022 at 10 pm UTC

MODULE 4

PRACTICAL APPLICATION 5.1

Practical Application Assignment 5.1: Will the Customer Accept the Coupon?

Published | Edit

Learning Outcome Addressed:

7. Apply exploratory data analysis, plotting, statistical summarization, and data visualization skills and techniques to a machine learning problem

Overview:

In this first practical application assignment of the program, you will seek to answer the question, "Will a customer accept the coupon?" The goal of this project is to use what you know about visualizations and probability distributions to distinguish between customers who accepted a driving coupon versus those that did not. Use the Practical Application 1 Jupyter Notebook to complete this assignment.

Data:

This data comes to us from the UCI Machine Learning repository and was collected via a survey on Amazon Mechanical Turk. The survey describes different driving scenarios, including the destination, current time, weather, passenger, etc., and then asks people whether they will accept the coupon if they are the driver. Answers given that the users will drive there "right away" or "later before the coupon expires" are labeled as "Y = 1", and answers "no, I do not want the coupon" are labeled as "Y = 0". There are five different types of coupons—less expensive restaurants (under $20), coffee houses, carry out and take away, bars, and more expensive restaurants ($20–$50).

EXERCISE (BUSINESS PROBLEM)

Our company sold an automotive product for over 20 years. However, for the last 5 years the monthly average profit has been constant and did not gain any significant growth since the number of sales are remain stagnant as well. The condition will remain the same in the future if we do not do something. We have a lot of customer leads that can be a potential buyer. However, with limited member of sales team, we don't have enough resource to approach more customer. It would be very inefficient and wasting a lot of resource to target all the leads. We want to be efficient instead of keep expanding the team, so we need another approach. With limited time and resources, we need to be able to quickly inspect and prioritize which customer is a potential buyer. We will also need to formally research on what makes them buy our products. By doing this, we can achieve higher or the same amount of profit with cheaper cost

In summary, our business problem is:

- We have stagnant profit because the number of sales is constant
- There are a lot of customer leads but we can't reach all of them
- We need to know which customer leads that should be prioritized
- We need lead scoring so that we can be efficient on targeting potential buyer

Source: https://rpubs.com/Argaadya/crispr_dm

EXERCISE (DATA)

Data are collected from the sales department in tabular format. The data consists of the past sales team interaction with the lead customer. The sales team keep record on whether the leads turn into purchase or refuse to buy the product, complete with the customer demographic information.
The collected data consists of 40,000 distinct customers with 14 variables. The description of each column/variable can be seen below:

- **flag** : Whether the customer has bought the target product or not
- **gender** : Gender of the customer
- **education** : Education background of customer
- **house_val** : Value of the residence the customer lives in
- **age** : Age of the customer by group
- **online** : Whether the customer had online shopping experience or not
- **customer_psy** : Variable describing consumer psychology based on the area of residence
- **marriage** : Marriage status of the customer
- **children** : Whether the customer has children or not
- **occupation** : Career information of the customer
- **mortgage** : Housing Loan Information of customers
- **house_own** : Whether the customer owns a house or not
- **region** : Information on the area in which the customer are located
- **fam_income** : Family income Information of the customer(A means the lowest, and L means the highest)

EXERCISE (DATA EXPLORATION AND QUALITY CHECK)

We need to check the quality of the data. For example, since many of the column/variable is categorical, we can check the summary of the data and see the number of customer of each categories. By doing this, we can also check whether there are any data that need to be cleansed or to be transformed. For example, we can check if there is a missing/empty values.

```
##  flag      gender              education       house_val          age
##  N:20000   F:16830                   :  741   Min.   :      0   1_Unk :6709
##  Y:20000   M:22019   0. <HS          : 3848   1st Qu.:  80657   2 <=25:2360
##            U: 1151   1. HS                                      
##                      2. So                                      
##                      3. Ba                                      
##                      4. Gr                                      
##
##  online     customer_psy
##  N:12681   B       :8197
##  Y:27319   C       :7830
##            E       :6650
##            F       :4058
##            G       :3951
##            D       :2353
##            (Other):6961
##   mortgage     house_owner
##  1Low :29848          : 337
##  2Med : 4803   Owner :2923
##  3High: 5349   Renter: 739
##
##
##
##
##
```

There are some interesting finding from the summary. For example, the gender column consists of 3 categories: F (Female), M (Male), and U (Unknown). The child column is similar, with additional value of U (Unknown) and 0 (zero) even though the column should only be Yes or No. The marriage and education column contain empty values. This is not surprising, since the sales team are not instructed to fulfill each column with pre-determined values. However, this means that the incoming data quality is not good and require future standardization in the future. This also show us that we need to cleanse and prepare the data before we do any analysis so that all relevant information can be captured.

EXERCISE (DATA CLEANSING)

Based on our finding, we will do the following process:

- Change missing/empty value in education, house_owner and marriage into explicit Unknown
- Make all *U* value in all categorical column into explicit Unknown
- Cleanse the age category by removing the index (1_Unkn into Unknown, 2_<=25 into <=25, etc.)
- Cleanse the mortgage category by removing the index

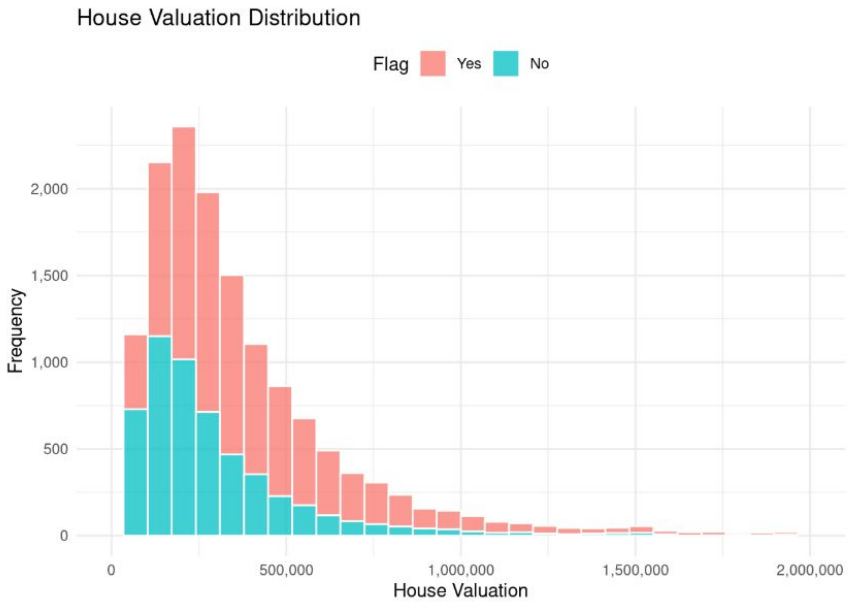
```
##  flag           gender             education       house_val
## No :20000   Female :16830   <HS          : 3848   Min.   :      0
## Yes:20000   Male   :22019   Bach         : 9267   1st Qu.:  80657
##             Unknown: 1151   Grad         : 5916   Median : 214872
##                             HS           : 8828   Mean   : 307214
##                             Some College:11400   3rd Qu.: 393762
##                             Unknown      :  741   Max.   :9999999
##
##      age       online       customer_psy     marriage         child
## <=25   :2360   No :12681   B      :8197   Married:20891   No     :13333
## >65    :4822   Yes:27319   C      :7830   Single : 5082   Unknown: 8655
## 26-35  :4984               E      :6650   Unknown:14027   Yes    :18012
## 36-45  :7115               F      :4058
## 46-55  :8103               G      :3951
## 56-65  :5907               D      :2353
## Unknown:6709               (Other):6961
##         occupation    mortgage      house_owner          region
## Blue Collar  : 6621   High: 5349   Owner  :29232   Midwest  : 8107
## Farm         :  329   Low :29848   Renter : 7391   Northeast: 7247
## Others       : 2006   Med : 4803   Unknown: 3377   Rest     :  245
## Professional :14936                                South    :15676
## Retired      : 4341                                West     : 8725
## Sales/Service:11767
##
##   fam_income
## E       : 8432
## F       : 6641
## D       : 4582
## G       : 4224
## C       : 2687
## H       : 2498
## (Other):10936
```

EXERCISE (EXPLORATORY DATA ANALYSIS)

Here we will do visualization to see whether there are any difference between customer who buy our product and who don't. To visualize a distribution, we can use histogram. The *x-axis* is the house valuation while the *y-axis* show the frequency or the number of customer with certain house valuation.

From the histogram, most of our customer has house valuation less than 2,500,000. Some customers are outlier and has house valuation greater than 2,500,000. Their frequency is low and they cannot be seen on the histogram. The distribution for people who buy and not buy are quite similar, therefore we cannot simply decide if a customer will buy our product based on their house valuation.

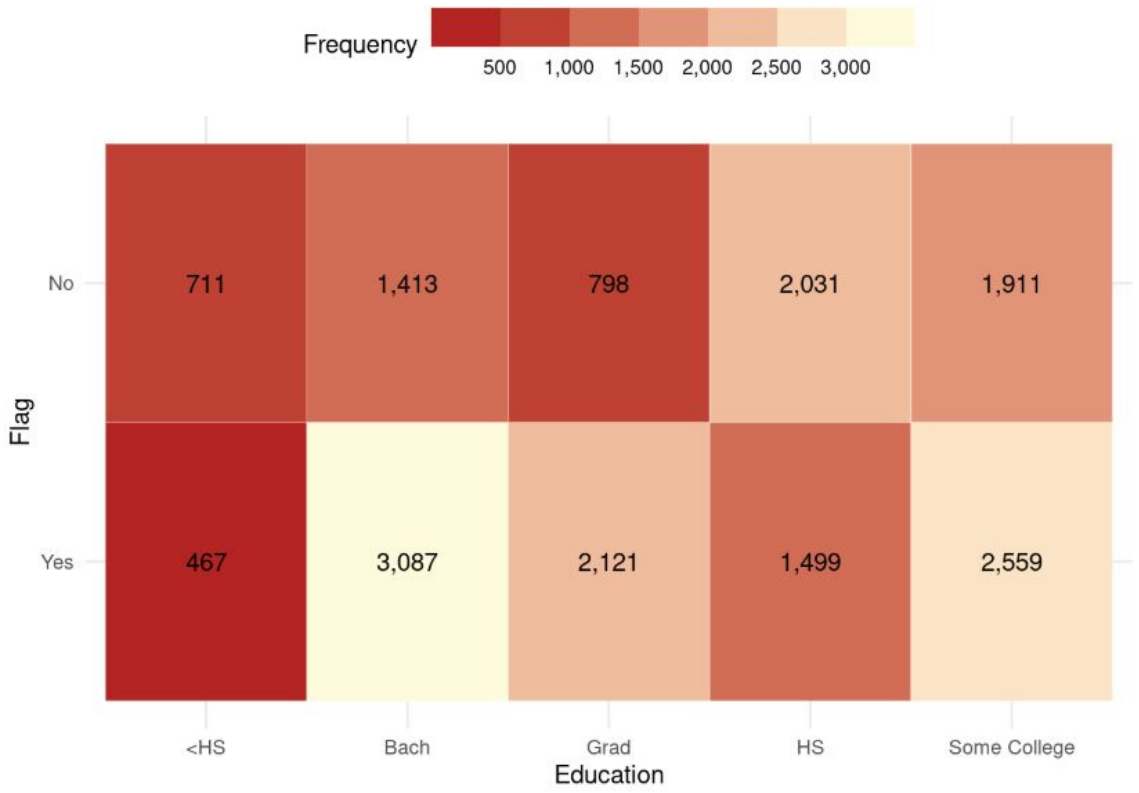
Without Outliers



EXERCISE (EXPLORATORY DATA ANALYSIS)

We will see if the education level can be a great indicator to decide if a customer has high probability to buy our product. The color of each block represent the frequency of people that fell in that category, with brighter color indicate higher frequency.

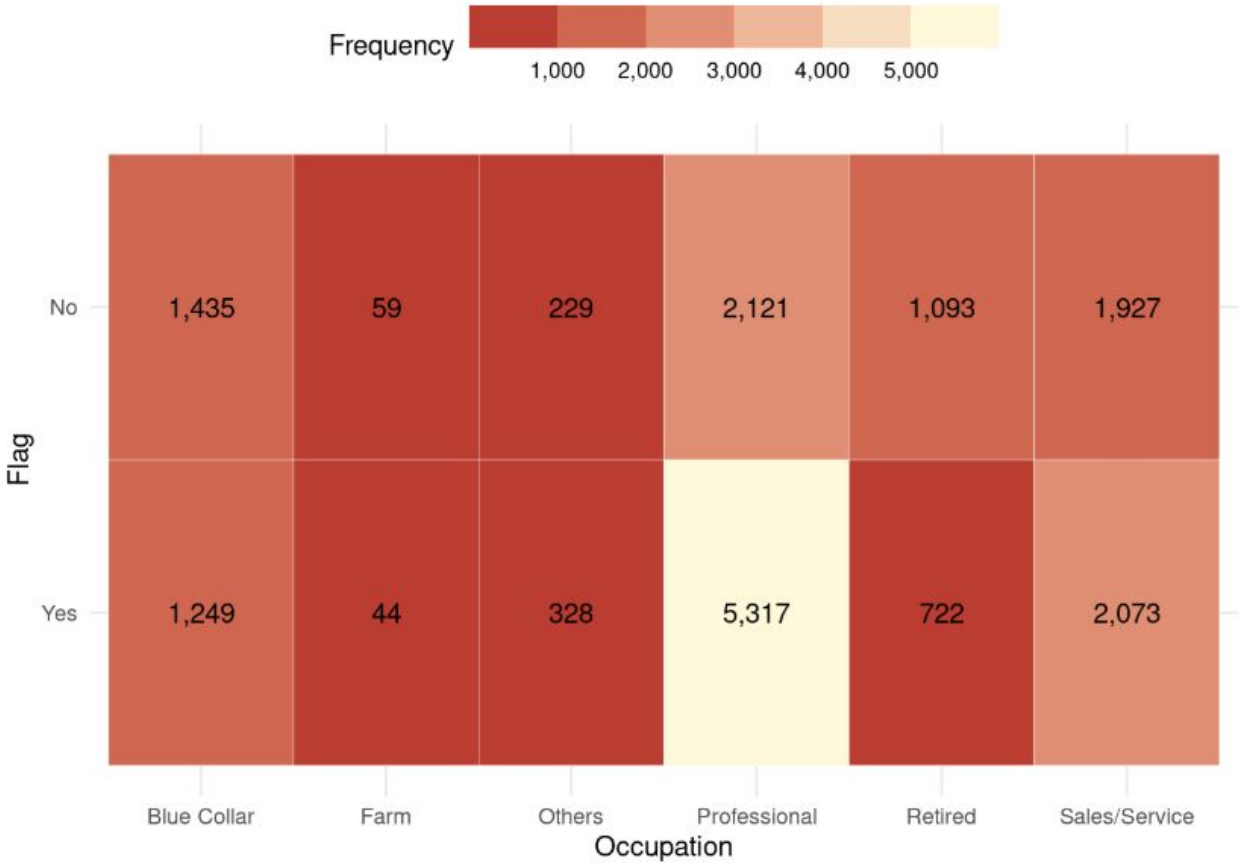
Based on the heatmap, people with higher education level (*Bach* and *Grad*) are more likely to buy our product. Therefore, education level may be a great indicator to check potential customer.



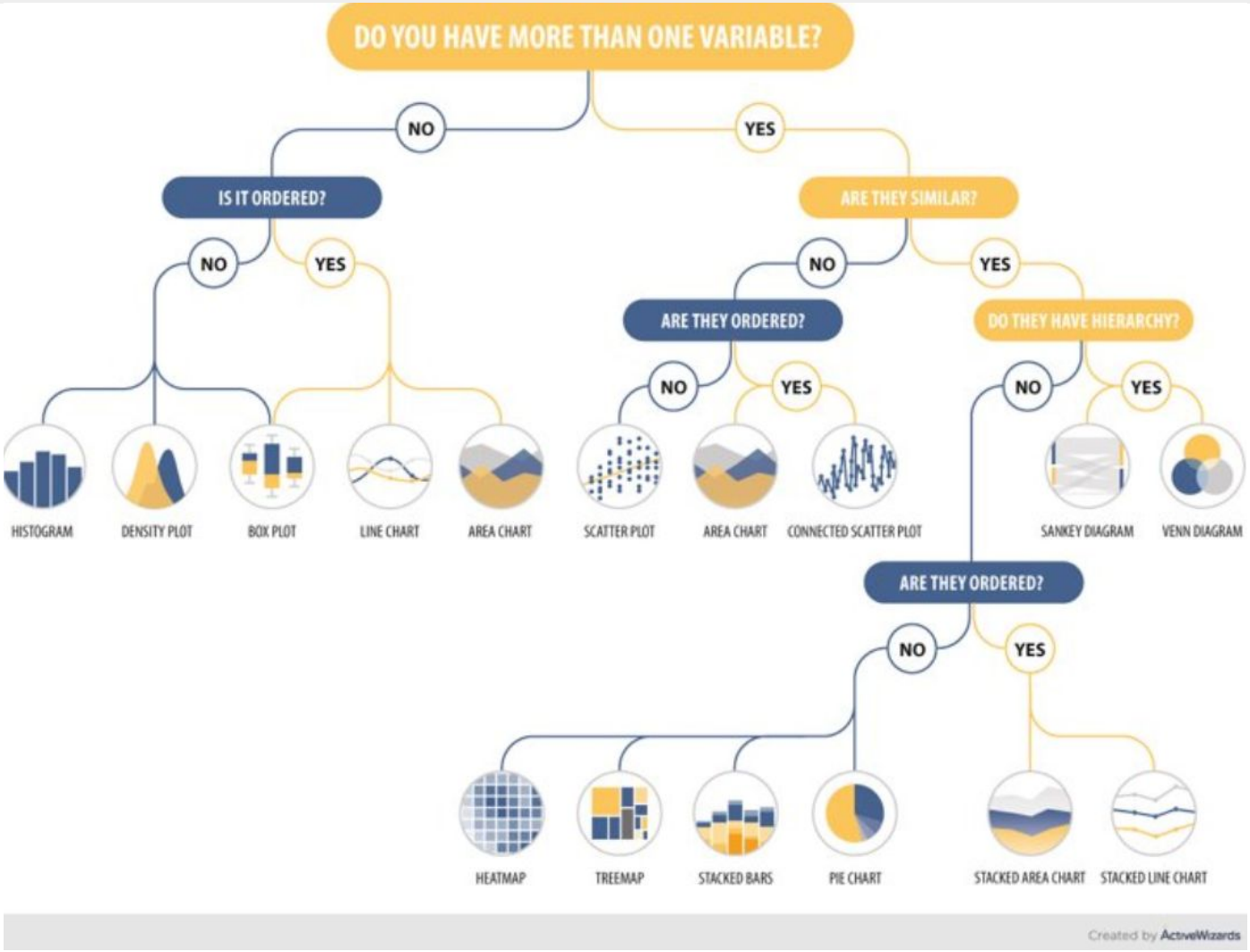
EXERCISE (EXPLORATORY DATA ANALYSIS)

We will do the same thing here with the occupation/job. The one that stands out is the professional occupation that has a very high frequency of people who buy our product.

We can keep doing this analysis with other variables/plots



DATA VISUALISATION FUNDAMENTALS



https://www.kdnuggets.com/2019/03/how-choose-right-chart-type.html

QUESTIONS?