Homework 8 Programming Task (10 points)

IST 5520 - Fall 2022, Chen

Name: __Ronald Adomako____

Data

The data file "UniversalBank.csv" contains a dataset of 5000 customers of the Universal Bank. Below is the description of columns in the dataset.

- Id: Customer ID
- Age: Customer's age in completed years
- Experience: #years of professional experience
- Income: Annual income of the customer ($000)
- ZIPCode: Home Address ZIP code.
- Family: Family size of the customer
- CCAvg: Avg. spending on credit cards per month ($000)
- Education: Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
- Mortgage: Value of house mortgage if any. ($000)
- Personal_Loan: Did this customer accept the personal loan offered in the last campaign?
- Securities_Account: Does the customer have a securities account with the bank?
- CD_Account: Does the customer have a certificate of deposit (CD) account with the bank?
- Online: Does the customer use internet banking facilities?
- CreditCard: Does the customer use a credit card issued by UniversalBank?

Note: Since Education should be categorical variable, you need to convert it into dummies.

Complete the following task and questions.

Programming Task: (6 points)

Conduct a multiple linear regression analysis. Regress Personal_Loan on other variables except Id and ZIP_Code.

Question 1: (2 points)

According to the OLS, interpret the effect of Education on Personal_Loan? Does the effect statistically significant? Since the Personal_Loan is a dummy variable, you can interpret the coefficient as the change of probability. Explain the effect using business language (people with common business background can understand).

Your Answer:

On average, no, Education does not have significant effect on predicting personal after having at least one degree. I showed this by parsing out the three education types into dummy variables. In the first case, having an undergraduate degree is the only example where the P>|t| is under 5%, with the other types of degree present, suggesting that more degrees does not help one get a loan. It would be helpful to see non-degrees in this dataset to further compare, but it is clear the first degree is significant in whether one obtains a loan.

Question 2: (2 points)

According to the OLS, interpret the effect of Age on Personal_Loan? Does the effect statistically significant? Explain the effect using business language (people with common business background can understand).
Your Answer:

Age is statistically significant on whether one gets a personal loan, because the p-value is less than 5% (suggesting a high confidence interval). However, the impact of the Age variable is low: with 0.79%, one would have to at least 126 years old to guarantee acceptance of a loan or at least 63 years old to have a 50% chance or receiving a loan.

Extra Credit

Doing a regression analysis is a good step to identifying variables for an iterative dimensional analysis (most likely achievable in one step). I noticed the OLS suffered computationally when trying to pre-scale the data like we do for PCA.

From this practice, I can see that if I cared about keeping the original components and not wanting to transform variables into a different coordinate space, I could run regression analysis to retain the significant variables iteratively until I got a result where I didn't have a high conditional score. This would help me identify the most significant variables.

If I cared about just predicting the whether the loan acceptance would be correctly predicted, then I could do a classification analysis employing PCA and scaling to find a model that clusters between loan accepted and loan rejected - Cluster Analysis. This may retain some of the lower significant variables and forego the insignificant ones although they wouldn't be explicitly mapped one-to-one since PCA transforms the coordinate space. The dimension reduction would be a hint of the number of significant variable - noting that all variables have a probability of significance or insignificance - partial significance. Classification analysis with PCA optimizes the weights of these variables while minimizing effects of multicollinearity otherwise the model wouldn't predict well on new data!

Submission:

Submit: (1) your jupyter notebook with results, and (2) this task document with answers.