Data Mining and Business Analytics

**Homework 3**

In this homework you will do exercises in KNN and Cross-validation. Please follow the instruction given in the homework and answer the question.

**Submission format:**

1. Please include your answers and the corresponding codes/plots/outputs in **one single word doc or pdf file**. If you are using R markdown that is great, you can submit the pdf output file. If you are not using R markdown, please include the corresponding codes/plots/outputs by screenshots or copy&paste.
2. Please also submit your **R script or R markdown file**. But please note that TA is not responsible to run your code to obtain any output to support your answers.

**Notes:**

1. This homework is a group assignment.
2. If you encounter any challenges, before coming to TA or me:
   1. Make sure you go through the “R and R helps.pdf” (posted under **R HELPS** Module on Canvas). Please pay special attention to “Few Notes Regarding R” and “Few Tips When R Is Not Working”.
   2. Ask Online using the following links
      1. Stackoverflow - <http://stackoverflow.com/> - coding and language related questions
      2. Statexchange - <http://stats.stackexchange.com/> - statistical related questions

**Additional Files on canvas for this homework:**

1. **credit\_data.csv**
2. **KNN calculation (15 PTS)**
3. Below is the dataset of 5 customers with their age, the amount of loan taken and if they had defaulted or not. Also, the standard deviation and mean values of their age and loan are given.

|  |  |  |
| --- | --- | --- |
| Age (in years) | Loan (in USD) | Default (Y-Defaulted, N- Not Defaulted) |
| 25 | 5,000 | Y |
| 40 | 4,000 | N |
| 45 | 6,000 | N |
| 30 | 8,000 | Y |
| 20 | 10,000 | N |

Standard Deviation of Age= 10.36

Mean of Age=32

Standard Deviation of Loan=2408.31

Mean of Loan=6600

Z-score normalization: Standardized value (X) = (X-Mean)/standard deviation

1. (3pts) Calculate the standardized age of each customer using z-score normalization.
2. (3pts) Calculate the standardized loan of each customer using z-score normalization.
3. (6pts) Suppose a person of age 35 requests for a loan amount $10,000. Calculate his/her standardized age and loan amount, and the Euclidian distance to all the other rows based on the standardized values.
4. (3pts) Based on k-nearest neighbor algorithm (k=3), what will be the predicted Default status of the new loan applicant mentioned in part c)?

**Answer:**

a) Calculating the standardized age and standardized loan of each customer.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age (in years) | Loan (in USD) | Standardized age | Standardized Loan | Default (Y-Defaulted, N- Not Defaulted) |
| 25 | 5,000 | -0.67568 | -0.66437 | Y |
| 40 | 4,000 | 0.772201 | -1.0796 | N |
| 45 | 6,000 | 1.254826 | -0.24914 | N |
| 30 | 8,000 | -0.19305 | 0.581321 | Y |
| 20 | 10,000 | -1.1583 | 1.411778 | N |

b) Calculating the standardized age and standardized loan of the new person and the Euclidian distance to all other rows

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Age (in years) | Loan (in USD) | Standardized age | Standardized Loan | Euclidean Distance | Default (Y-Defaulted, N- Not Defaulted) |
| 25 | 5,000 | -0.67568 | -0.66437 | 2.289560256 | Y |
| 40 | 4,000 | 0.772201 | -1.0796 | 2.537689868 | N |
| 45 | 6,000 | 1.254826 | -0.24914 | 1.921028511 | N |
| 30 | 8,000 | -0.19305 | 0.581321 | 0.960514256 | Y |
| 20 | 10,000 | -1.1583 | 1.411778 | 1.447876448 | N |
| 35 | 10000 | 0.289575 | 1.411778 | 0 | To be |
|  |  |  |  |  | classified |

c) Ranking the minimum Euclidian distance

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Age (in years) | Loan (in USD) | Standardized age | Standardized Loan | Euclidean Distance | Default (Y-Defaulted, N- Not Defaulted) | Rank minimum distance |
| 25 | 5,000 | -0.67568 | -0.66437 | 2.289560256 | Y | 4 |
| 40 | 4,000 | 0.772201 | -1.0796 | 2.537689868 | N | 5 |
| 45 | 6,000 | 1.254826 | -0.24914 | 1.921028511 | N | 3 |
| 30 | 8,000 | -0.19305 | 0.581321 | 0.960514256 | Y | 1 |
| 20 | 10,000 | -1.1583 | 1.411778 | 1.447876448 | N | 2 |
| 35 | 10000 | 0.289575 | 1.411778 | 0 |  |  |

d) Since k=3 and majority (2 out of 3) nearest neighbor of the new person have a Default status of N, the new person is classified as having a default Status N.

**For the next part, we will be using the bank loan default data (credit\_data.csv).**

**Dataset Description**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Column** | **Description** |
| 1 | months\_loan\_duration | Duration of the loan |
| 2 | credit\_history | Previous credit history |
| 3 | purpose | Purpose of the loan |
| 4 | amount | Amount borrowed |
| 5 | employment\_duration | Present employment since |
| 6 | percent\_of\_income | Installment rate in percentage of disposable income |
| 7 | years\_at\_residence | Present residence since |
| 8 | age | Age of the borrower in years |
| 9 | other\_credit | Any other credit |
| 10 | housing | Housing type of borrower |
| 11 | existing\_loans\_count | Number of existing credits at this bank |
| 12 | job | Job type |
| 13 | dependents | No, of dependents of the borrower |
| 14 | phone | If the borrower has a phone or not |
| 15 | default | Classifier-if the Borrower defaulted or not |

1. **KNN and Cross-validation (18 PTS)**

**Using the bank loan default data.**

1. (16 pts) Implement K nearest neighbor classification with 10-fold Cross-validation on the given dataset and plot the accuracy based on K=1 to 50. (Hint: use caret package to implement K nearest neighbor. Also use set.seed() function with seed=100)
2. (2 pts) What K value gives the highest accuracy?

Answer:

![Chart, line chart, scatter chart

Description automatically generated]()

Using the function which.max, which.max(knn\_credit$results$Accuracy) to find out the K with the highest accuracy.

From the above plot, k = 20 gives the highest accuracy.

1. **Refer back to part 5 of your homework 2 in which you have found a dataset. Feel free to change the dataset. (17 PTS)**
2. (2pts) Provide the name of the dataset and source. Each group only needs to focus on **one** dataset.
3. (5pts) What type of companies/institutions (or other entities) would be able to gain insights (e..g, competitive advantage) from utilizing this data? You could come up with a specific name of a company or broad industry level description. Explain why.
4. (5pts) What would be an interesting supervised learning task associated with this data for gaining these insights? Identify the y variable of interest and justify why y is interesting.
5. (5pts) Is there any additional data you could obtain that will make the current dataset more powerful? How would you go about obtaining that auxiliary data? It is okay if this additional data is impossible to get assess to. You can just describe hypothetically what information would be helpfup to argument the current dataset.

All relevant answers are given full marks.