

|  |
| --- |
| PERFORMANCE ASSESSMENT OF  DATA MINING II  TASK2  D212  BY KOFFI M. GANU |
| MSDA MAY 2022 AT WGU  INSTRUCTOR: KESSELLY KAMARA  Phone  Email |

**INTRODUCTION**

People often say in business: “It takes a month to find a customer but a second to lose one.” The primary goal of any company is to maintain its customers as long as possible. It is with this in mind that, especially with the advent of the use of data, these companies hire data specialists (Data analysts, Data scientists) for strategies and recommendations for the improvement of their company’s services. In the rest of our project, we will use the principal component analysis (PCA) method to analyze the data set provided to identify the principal variable, ultimately allowing strategic decision-making.

**PART I/**

A1-

What is the profile of the customers who are most likely to churn?

A2-

Our goal is to identify the key factors that contribute to customer churn. By applying the principal component analysis (PCA), we can determine the most significant variables contributing to churn and gain insights into customer behavior patterns.

**PART II/**

B1

The principal component analysis is a widely used technique in data analysis and dimensionality reduction. This technique transforms a high-dimensional dataset into lower dimensional representation while retaining the most important information. To perform the PCA analysis, we will follow thesetheses steps:

* Data scaling.

PCA begins by scaling the dataset to ensure that all variables have a mean of zero and a standard deviation of one.

* Covariance matrix calculationcalculations:

The covariance matrix provides information about the relationships between pairs of variables.

* Eigendecomposition:

Eigendecomposition is a mathematical process that decomposes a matrix into a set of eigenvectors and eigenvalues. The eigenvectors represent the directions or principal components, and the eigenvalues indicate their respective importance.

* Selection of principal components:

The eigenvectors are ranked based on their corresponding eigenvalues, with the eigenvectors associated with the highest eigenvalues being the first principal component, the one with the second highest eigenvalues being the second, and so on.

* Variance explained:

The eigenvalues associated with the principal components represent the proportion of variance explained by each component. This information helps in determining the importance of each principal component.

B2-

One assumption of PCA is that the variables should have a linear relationship. Indeed, PCA assumes that their relationship can be represented by a straight line in higher dimensions. If the variables exhibit non-linear relationships, there may be better techniques than PCA, as it may not capture the underlying patterns effectively.

**PART III/**

C1-

The continuous dataset variable used to perform the PCA are:

A close-up of a white background

Description automatically generated

C2-

To perform PCA, it is essential to standardize the variables to ensure that each variable contributes equally to the PCA process. For our project, we have used standarscaler package, and here is the result.

A computer screen shot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

**PART IV/**

D1-

We will use the eigenvectors obtained from the analysis to determine the matrix of all the principal components. Each eigenvector represents a principal component stacked together to form the matrix of principal components.

A table of numbers and symbols

Description automatically generated

A table of numbers and letters

Description automatically generated

A table of numbers with text

Description automatically generated

A screenshot of a computer code

Description automatically generated

D2-

To identify the total number of principal components to retain in PCA, we will use the Kaiser Criterion method. This method is a statistical rule that suggests retaining only the principal components with eigenvalues greater than 1. According to the Kaiser Criterion, only components with eigenvalues

greater than one are considered significant, as they explain more variance than a single original variable. Therefore, we will retain nine principal components.

A graph with a red line and a blue line

Description automatically generated

D3-

The variance of each principal component can be identified using the eigenvalues obtained from the PCA. The eigenvalues represent the amount of variance explained by each principal component.

A graph of a person walking up a staircase

Description automatically generated

D4-

The total variance gives us the total variance captured by all the principals retained. This value represents the proportion of the total variance in the original data explained by the principal components. According to our analysis, the total variance of our nine components retained is 80%. So the total variance is 0.8

D5-

PCA is a dimensionality reduction technique. It creates new variables using linear combinations of old variables. PCA is designed to create independent variables and manages to tell us how important each of these new variables is. This importance helps us to choose how many variables we will use. To pick the number of PCA, we will create a scree plot showing the eigenvalue for each component and use the Kaiser Criterion methodology. According to Kaiser Criterion, If the eigenvalue is greater than 1, the principal component is better at explaining the variance in the data set. In our case, we will keep the first nine PCA. In addition, we will see on the” Cum of variation explained” graph PC1, PC2, PC3, PC4, PC5, PC6, PC7, and PC8 make up to 80% of the variance. That confirms that the first nine components were deemed most important to the analysis.

**PART V/**

E- PANOPTO VIDEO

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=7126a859-c00f-4708-9d95-b05e011474ff

F-

No in-text references were used.

G-

We used professional communication in the presentation of the submission.