**CAN PRINCIPAL COMPONENT ANALYSIS BE CREATED USING A SET OF CUSTOMER ATTRIBUTES?**

**TASK 2**

**D212**

**STUDENT NAME: IBRAHIM SULEIMAN**

**DATA ANALYTICS**

**STUDENT NUMBER: 001429984**

**COLLEGE OF INFORMATION TECHNOLOGY, WESTERN GOVERNORS UNIVERSITY**

**NAME OF INSTRUCTOR: KEIONA MIDDLETON**

**NOVEMBER 25th, 2022**

**Part 1: Research Question**

Section A: Description of the Report

Section A1: Can principal component analysis be created using a set of customer attributes?

This is relevant to a real-world organizational situation and would be answered using **principal component analysis.**

Dimensionality reduction is the process of lowering the number of variables used as inputs into a predictive model.

Principal Component Analysis, or PCA for short, is a prominent method for dimensionality reduction in machine learning.

Section A2: Goal of Analysis

The objective of this analysis is to use PCA for dimensional reduction(elimination) in order to find the factors that influence turnover in the telecom industry. In order to project high-dimensional data onto a new subspace with the same number of dimensions or fewer than the original one, this seeks to identify the directions of maximum variation in the data.

Dimensionality reduction is the process of lowering the number of variables used as inputs into a predictive model. A model with fewer input variables may perform better when making decisions based on fresh data (Jason Brownlee. 2020).

Part II: Method Justification

Section B1: How PCA Analyzes Churn Dataset

PCA will help us identify patterns based on the correlation between the traits in the churn dataset. In a word, PCA projects high-dimensional data onto a new subspace with the same number of dimensions as the original subspace in an effort to identify the directions of maximum variance.

In the end, this will enable improved business and strategic decision-making by significantly assisting in the reduction of enormous datasets, the discovery of the correlations between the variables, and the identification of the key variables of our clients attributes.

Section B2: Assumption

Principal components analysis (PCA), in contrast to factor analysis, makes the assumption that there is no unique variance and that the total variance is equal to the common variance. Examine the distinction between common and unique variance..

Part III: Data Preparation

Data Preparation Steps

Prior to conducting the analysis, the data must be available. The first step is to ensure that there are no blank columns in any of the columns. Making sure there are no duplicates of any of the data in the columns should come next. In addition, we want to ensure that there are no duplicate columns or rows, so we'll verify that as well (False).

The dataset includes certain variables that were found to be worthless for the logistic analysis, such as customer demographics that cannot be changed and are connected to the interaction and location of the consumer, thus those columns should be deleted. In order to give a clearer understanding and determination of applicable factors, the survey columns also need to be renamed (Peter Grant, 2019).

Section C1: Identified Continuous dataset variables

Graphical user interface, text, application

Description automatically generated

Diagram

Description automatically generated

Section C2: Standardized continuous dataset variables

Standardizing the data is always recommended before using PCA or any other machine learning technique. The most popular scalar for this is the Standard Scalar. Sklearn already includes a Standard Scalar. Therefore, we will now use Standard Scalar to standardize the feature set and store the scaled feature set in a pandas data frame. The data are scaled using StandardScaler, which sets the data's mean at 0 and standard deviation at 1.

Table

Description automatically generated with low confidence

Table

Description automatically generated

Copy of the Cleaned Dataset

The cleaned dataset was provided as csv file in the submission named;

We exported our prepared dataset as

df2.to\_csv('prepared\_d212task2.csv', index = False)

Part IV: Analysis

Section D1: The matrix of all the principal components

Application

Description automatically generated

Section D2: Identified the total number of principal components using kaiser criterions and a screenshot of the spree plot

Chart, line chart

Description automatically generated

On the x-axis is the principal component, and on the y-axis is the percentage of variation explained by each principal component. Using the Kaiser's criterion, factors with eigenvalues larger than 1 are retained. Therefore, the Principal components greater than 1 range from 0 to 6 from the above plot. We will keep 1 to 6 due to the fact that they contain eigen value greater than 1 and we will get rid of the rest (pc7, pc8, pc9, pc10, pc11).

This is predicated on the idea that it isn't statistically plausible to keep a component that speaks for less variance than a single original variable (Mumford Karen, 2003).

Section D3: Identified the variance of each of the principal component

Principal Component Analysis (PCA) is another tool we use to minimize the dimension of features by generating new features that retain the majority of the original data's variance. The value of the parameter n components, which represents the proportion of features in the final dataset.

It aids in the removal of linked characteristics when the principal components are reduced. We might be able to plot and visualize data more precisely if we reduce its dimensions.

Graphical user interface, text, application

Description automatically generatedThe numbers of principal components were reduced down from 11 to 6, we are keeping PC1 through PC6.

Based on the Kaiser’s rule plot the illustration which appears that PC1 = 16.63%, PC2 = 10.28% , PC3= 8.82%, PC4 = 8.62%, PC5 = 8.50%, PC6 = 8.37%, the eigen value greater than equals 1 and that is why they were selected.

Section D4: Identified the total variance captured by the principal components

Graphical user interface, text, application

Description automatically generated

About 93.55% of the total variation is covered by the principal components.

PC1 = 16.63%, PC2 = 10.28% , PC3= 8.82%, PC4 = 8.62%, PC5 = 8.50%, PC6 = 8.37%, the total variance of the PC we decide to keep = 61.22 % which gives us a feasible total variance feasible for further analysis. We can easily preserve 61.22% of the entire variance by utilizing just the six principal components, therefore we can comfortably say that these principal component fulfill the kaisar rule of greater than 1.

Section D5: Result summary

We observe:

PC1 = 16.63%, PC2 = 10.28%, PC3= 8.82%, PC4 = 8.62%, PC5 = 8.50%, PC6 = 8.37%.

Total PCAs Variance = 61.22%.

(‘Lat’)

the latitude location variable shows that PC1 make up about 16.63% variance of the dataset.

('Population')

Based on census statistics, the population that is PC2 represents the mile radius of the consumer.. This PC make up of about 10.28%, of the variance.

('Children')

Number of kids in the customer's home as stated in the registration data (may not be children of customer) as PC3 which makes up about 8.82%, of the variance.

('Age') the age variable clearly shows

Age of customer as reported in sign-up information illustrates that PC4 make up about 8.62% variance.

('Income')

Variable PC5 represents the annual income of the customer (or billed person) as reported at the time of enrolment. Its variance is 8.50%.

('Outage\_sec\_perweek') This variable represents the typical weekly duration of system outages in the customer's area. The percentage of PC6 is 8.37%.

According to an analysis of the Eigin values revealed in these PCAs, these are the most crucial criterion when applying Kaisar's Criteria to determine the consumers' qualities. PCA can generally be applied when the variables are strongly correlated, but it may be less useful when there is a weak connection.

Part V: Attachments

Copy of the dataset and codes were added in submission.

Sources

Kumar .B. (2022) *Eigen Factors Analysis.*<https://pythonguides.com/python-scipy-eigenvalues/>

Keren M. (2003). *Factors Retention in Exploratory Factor Analysis: A Comparison of Alternative Methods.* [https*://*Eric.ed.gov/?id=ED476430](https://Eric.ed.gov/?id=ED476430)

***How to reduce dimensionality using PCA in Python?***

<https://www.projectpro.io/recipes/reduce-dimentionality-using-pca-in-python>

Jason, B. (2020). **Principal Component Analysis for Dimensionality Reduction in Python.**

<https://machinelearningmastery.com/principal-components-analysis-for-dimensionality-reduction-in-python/>

Grant, P. (2019). Understanding Multiple Regression; The fundamental basis behind this commonly used algorithm.

Medium.<https://towardsdatascience.com/understanding-multiple-regression-249b16bde83e>

Evgeniy R.(2020). 5 Stages of Data Preprocessing for K-means clustering

Medium. <https://medium.com/@evgen.ryzhkov/5-stages-of-data-preprocessing-for-k-means-clustering-b755426f9932>

Michael G. (2018). Understanding Boxplots.

Medium. <https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51>

Mr D. (2018). K-Mean clustering for Wine Quality Data

<https://www.kaggle.com/code/digvijaysingh16/k-mean-clustering-for-wine-quality-data>

Neelam T. (2020). What is K-means Clustering in Machine Learning?

<https://www.analyticssteps.com/blogs/what-k-means-clustering-machine-learning>

Andrea T.(2016). Introduction to K-means Clustering

<https://blogs.oracle.com/ai-and-datascience/post/introduction-to-k-means-clustering>

How to Swap Between Colorblind and Colorful Dashboards. (2016, July 5). YouTube.

<https://www.youtube.com/watch?v=3iNl7KMK8pM>

Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. Wiley .(Chapter 1, Chapters 3-5, Chapters 7 - 8)

<https://ebookcentral.proquest.com/lib/westerngovernorsebooks/reader.action?docID=4187267&ppg=1>

Knaflic, C. N. (2015). storytelling with data: A data visualization guide for business professionals.

Wiley .

(Chapter 1, Chapters 3-5, Chapters 7 - 8)

<https://ebookcentral.proquest.com/lib/westerngovernors-ebooks/reader.action?docID=4187267&ppg=1>