Data Analysis - Customer Retention Project

1. Introduction

The project has 71 features, the aim is to reduce the features and take only the ones that are important for building a model. Too many features can impact the accuracy of the model.

I have also tried to extract correlations between different features with correlation martic and corr() method.

I have used the PCA method as a method for feature reduction to process the data to process it for machine learning

2. Problem Definition

The purpose is to highlight the key correlations and pick the important features out of the 71 features given in the data using the best methods for data analysis

3. Experimental Evaluation

3.1 Methodology

- Used power transform method to transform he data and remove outliers and skewness.
- Used the ‘Yeo Johnson’ method
- Used standard scalar method to scale the data and remove any outliers.
- We would be using this step for fruitful feature reduction as there are 71 features.
- We have used Principal Component Analysis (PCA)as a feature reduction method
- After treating the data, there are only 25 features that are required to handle data out of 71 features

3.2 Results

Using the correlation matrix and corelation method by sort function to arrange correlations between 2 variables in ascending order :

High correlation of 85% between **19 Information on similar product to the one highlighted is important for product comparison** & **22 Ease of navigation in website**

High correlation of 86% between **19 Information on similar product to the one highlighted is important for product comparison** & **22 Ease of navigation in website**

High correlation of 90% between **21 All relevant information on listed products must be stated clearly** & **The content on the website must be easy to read and understand**

High correlation of 90% between **21 All relevant information on listed products must be stated clearly** & **38 User satisfaction cannot exist without trust**

High correlation between **18 The content on the website must be easy to read and understand** and **21 All relevant information on listed products must be stated clearly -** 90%

High correlation between **26 Trust that the online retail store will fulfill its part of the transaction at the stipulated time** & **23 Loading and processing speed** with 84%

High correlation between **24 User friendly Interface of the website** & **25 Convenient Payment methods** with 90%

High correlation between **23 Loading and processing speed** & **26 Trust that the online retail store will fulfill its part of the transaction at the stipulated time** - 84%

High correlation between **23 Loading and processing speed** & **24 User friendly Interface of the website** - 82%

High correlation between **38 User satisfaction cannot exist without trust** & **27 Empathy (readiness to assist with queries) towards the customers** - 85%

High correlation between **31 Enjoyment is derived from shopping online** & **30 Online shopping gives monetary benefit and discounts** - 86%

High correlation between **30 Online shopping gives monetary benefit and discounts** & **31 Enjoyment is derived from shopping online -** 86%

High correlation between **35 Displaying quality Information on the website improves satisfaction of customers** & **32 Shopping online is convenient and flexible** - 82%

High correlation between **32 Shopping online is convenient and flexible & 35 Displaying quality Information on the website improves satisfaction of customers -** 82%

High correlation between **18 The content on the website must be easy to read and understand & 38 User satisfaction cannot exist without trust-** 89%

High correlation between **27 Empathy (readiness to assist with queries) towards the customers &** 38 User satisfaction cannot exist without trust - 85%

High correlation between **21 All relevant information on listed products must be stated clearly** & **38 User satisfaction cannot exist without trust** - 89%

High correlation between **18 The content on the website must be easy to read and understand** & **38 User satisfaction cannot exist without trust** - 89%

High correlation between **Availability of several payment options** & **Easy to use website or application** - 83%
High correlation between **Complete, relevant description information of products & Easy to use website or application -** 86%

High correlation between **Availability of several payment options & Complete, relevant description information of products -** 80%
High correlation between **Easy to use website or application** & **Complete, relevant description information of products** - 86%

High correlation between **Perceived Trustworthiness** & **Reliability of the website or application** - 93%

High correlation between **Availability of several payment options & Complete, relevant description information of products** - 83%

High correlation between **Availability of several payment options** & **Easy to use website or application** - 86%

4. Future Work

I would find a more convenient and less time consuming mthod more to deal with 71 features and find their correlations.

5. Conclusion

The final 25 features extracted out of 71 original features by the PCA method have no skewness negligible outliers