Zakaria Bouzeya 500761082

Mel van der Meij 500843758

Semen Mokrov 500954045

Teacher: Diptish Dey

Business Statistics

|  |  |
| --- | --- |
| Business statistics  Sub assignment 1 | Abstract  [Trek de aandacht van uw lezer met een veelzeggend citaat uit het document of gebruik deze ruimte om een belangrijk punt te benadrukken. Sleep dit tekstvak als u het ergens anders op de pagina wilt plaatsen.]  Zakaria Bouzeya  [Cursustitel] |

Main research question Tao Yin:

What segments are there for the company to anticipate on?

Sub question 1

*Which distinct user segments exist?*

Conclusion:

As a result of the analysis, there are 3 main groups of buyers who redirect their goods to another address:

1. Ordinary customers who participate only in promotions and have a small profit (cluster 2).

2. Ordinary customers who bring a small profit for the company and buy less than the average. (cluster 0)

3. Since the model was able to identify ordinary buyers, it becomes logical to assume that the buyers marked with cluster 1 are drop shippers who do not cause a negative effect to the company to a greater extent than cluster 2 does. However, they still cause significant losses to the company due to the high volume of purchases and profit below 0. Since there are fewer of them than Group 2, you don't have to single them out as a serious threat.

*Which distinct item feature clusters exist?*

The elbow method analysis indicates that K=5 is an optimal number of clusters for segmenting the dataset, as the inertia decreases significantly up to this point before leveling off. This suggests that five clusters can effectively capture the underlying structure in the data without overfitting.

Upon applying K-means clustering to the PCA-transformed features, the scatter plot reveals distinct groupings along the two principal components. The first two principal components account for a cumulative explained variance of approximately 99.5%, capturing nearly all the variability in the dataset. This indicates that the dimensionality reduction was successful, simplifying the visualization while retaining most of the information.

However, a key limitation in our analysis is that the features in the item\_features dataset are not clearly defined. Due to this lack of definition, it is challenging to interpret the clusters in a meaningful way or derive actionable business insights. Without a clear understanding of what each feature represents, it is not possible to link the clusters to specific customer behaviors or product attributes. As a result, while the clustering algorithm successfully identifies patterns in the data, these patterns cannot be translated into concrete business strategies or targeted actions.

To enhance the value of this analysis, it would be crucial to obtain a clear definition of the features in the item dataset. This would allow for more meaningful interpretation of the clusters and the development of actionable insights, such as customer segmentation or personalized marketing strategies