Davide Quaglio 597201
Margarita Tavkazakova 624750
Jinfei Chen 621195
Panagiotis Terzopoulos 629456

Task 1 Segmentation

1.1

To effectively segment consumers and identify those who align with Intel's marketing goals, a survey was conducted to gather data on participants' preferences for various product characteristics. The survey results were used as the basis for segmentation, focusing specifically on participants' rated attributes about the importance of a product. In addition to product-related attributes, participants' willingness to pay was also included as a variable in the clustering process. Demographic and sociological attributes were also collected, but were not utilised as the primary variables for segmentation. Instead, these attributes were used to provide additional context and descriptive information about the consumers in the resulting segments.

In order to determine the most suitable clustering model for the given dataset, various criteria must be examined. To begin, the main characteristics of each model had to been considered. K-means is well-suited for large datasets, providing general patterns and broad trends. In contrast, hierarchical clustering is more appropriate for breaking large clusters into smaller sub-clusters and should be used when different levels of granularity are needed.

Secondly, to determine the optimal number of clusters for our analysis, various measures were employed, including the Elbow method and Silhouette analysis for k-means, as well as the Calinski-Harabasz index for HC-Ward. The results of the Silhouette analysis indicated that a model with three clusters provided a better match to the observations than other k-values. The Elbow curve shows instead more ambiguity, beginning to level off after k=3, with no notable decrease observed for k=4 and k=5. As a result, three clustering solutions (3, 4, and 5 clusters) were considered. The Calinski-Harabasz index yielded a maximum score at k=3, suggesting that the generation of three clusters resulted in a better separation of the data points. However, the smoothness of the line made it difficult to determine the most appropriate number of clusters, as the index values for 4 and 5 clusters were not significantly different in terms of variance. Also for this model we will produce 3 different solutions (3 ,4 and 5 clusters). Third, to ensure that the clusters generated by both models were visually distinct and did not overlap excessively, a detailed examination of the clusters was also conducted. Figure 1 illustrates the clusters generated by both models.

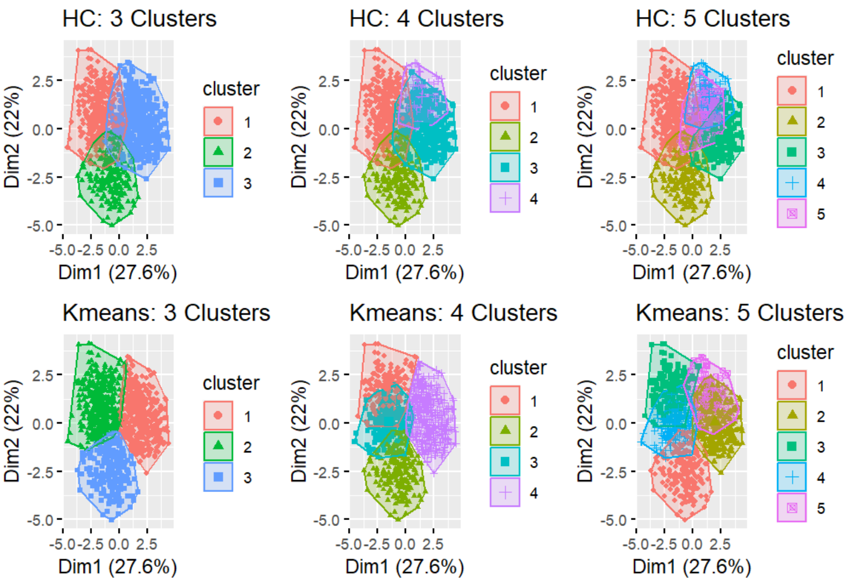

Figure 1 Clusters generated by HC and Kmeans

Davide Quaglio 597201
Margarita Tavkazakova 624750
Jinfei Chen 621195
Panagiotis Terzopoulos 629456

In the case of K-means with 3 clusters, the groups are clearly differentiated and the cluster-size is almost balanced. In contrast, the solutions with 4 and 5 clusters, the overlapping is visible and the groups are unbalanced, with one cluster containing 400 participants and another with only 185.

On the other hand, for all the solutions in the hierarchical model, a significant portion of clusters overlap, indicating poor segmentation. Finally, based on the reasons explained above, we can conclude by opting for the k-means model with 3 clusters, as it presents the best balance between the criteria analysed.

1.2

According to the means of the participants' rated attributes depicted in Table 1, we propose three segments:

- Segment 1: Health enthusiasts.
    - Largest segment (42%).
    - Participants in this cluster value many attributes of the product which is shown by the fact that 6 different attributes have a rating higher than 5. The highest rating of the product is about well-being (5.77) and the second highest is about style (5.29).
    - Participants in this cluster are mostly occupied in the fields of Finance (16%) and Advertisement (14.68%).
    - Participants in this cluster mostly use Facebook/Instagram (88.19%) and rather prefer to watch Amazon Prime (73.53%) or read magazines/news (63.13%) than watching TV (58.07%).
    - Participants in this cluster are the youngest compared to the other 2 clusters (30.66), mostly females (66.02%), hold the highest level of education (1.42) and are willing to pay 215.42 for the smartwatch.
- Segment 2: Apathetics.
    - Second largest segment (37%).
    - Participants in this cluster didn't rate highly over the importance ratings of the smartwatch, the only feature they highly value over the other two segments is the ability of the smartwatch to take a photo (3.66). Despite that, the highest value for them is the creative communication of a smartwatch (4.23).
    - Participants in this cluster are mostly occupied in the fields of sales (14%), education (13.74%) and retail (13.74%).
    - Participants in this cluster mostly use Facebook/Instagram (70%) over the other options and mostly watch TV (95.95%).
    - Participants in this cluster are the least educated, compared to the other clusters, since the highest level of education is 1.18, while also their willingness to pay for the product (189.35) and their household earnings (2.70) were found to be the lowest of the 3 clusters.
- Segment 3: Constant communicators.
    - Smallest segment (21%)
    - Participants in this cluster value highly the abilities of the smartwatch that offers up-to-minute smart traffic updates (5.56) and constant communication with family, friends and colleagues (5.48).
    - Participants in this cluster are mostly self-employed (22.89%) or occupied in the field of Sales (23%).

- Participants in this cluster prefer to use Podcasts or Radio (71.10%), Twitter (61.21%) and YouTube (61.68%) over the other options. Also, they mostly watch TV (91.58%).
- Participants in these clusters are the oldest (39.11), their household earns the most income (3.74) compared to the other segments, almost half of them (49%) receive technological devices from their employer and they are willing to pay the highest price for the smartwatch compared to the other clusters (248.69).

Variables	1	2	3
size	0.420	0.370	0.210
Imp_Innov	4.480	3.690	4.060
Imp_ConstCom	5.030	3.850	5.480
Imp_CreatCom	5.230	4.230	2.430
Imp_TimelyInf	4.400	3.440	5.560
Imp_SaveMT	4.010	3.600	2.110
Imp_SaveML	5.080	2.890	3.740
Imp_TaskMgm	4.730	3.210	4.850
Imp_DeviceSt	4.330	2.580	5.070
Imp_Photo	3.290	3.660	2.050
Imp_Wellness	5.770	3.170	3.720
Imp_Athlete	5.110	3.090	2.560
Imp_Style	5.290	3.640	3.510
WTP	215.420	189.350	248.690
iPhone	0.690	0.350	0.560
CompBuy	0.150	0.090	0.490
Occup_Health	0.110	0.030	0.030
Occup_Finc	0.160	0.090	0.130
Occup_Sales	0.100	0.140	0.230

Variables	1	2	3
Occup_Cons	0.024	0.105	0.117
Occup_Advt	0.147	0.102	0.023
Occup_Edu	0.067	0.137	0.028
Occup_Eng	0.051	0.027	0.098
Occup_Tech	0.120	0.100	0.065
Occup_Retail	0.058	0.137	0.014
Occup_SMB	0.043	0.084	0.229
FB_Insta	0.882	0.701	0.547
Twit	0.448	0.469	0.612
Snap	0.622	0.189	0.173
YouTube	0.607	0.493	0.617
Pod_radio	0.554	0.550	0.710
TV	0.581	0.960	0.916
NewsP	0.631	0.555	0.729
AmznP	0.725	0.410	0.519
Age	30.660	38.871	39.112
Female	0.660	0.523	0.458
Degree	1.427	1.181	1.411
Income	3.581	2.709	3.743

Table 1: The means of the participants' rated attributes

1.3

Upon consideration, we have concluded that the meaningful criteria Intel should focus on are the size of each cluster, the amount of money customers is willing to pay for a smartwatch and the annual household income that they earn. In order to rate the criteria, it seemed more accurate to rate them based on their importance (in a 5-point scale) as attractive characteristics for Intel to base on and draw customers from. Having that in mind, the parameters, size and willingness of the customer to pay, have both been rated with an importance of 40% and the annual household income is rated with an importance of 20%. In order to decide on the segment Intel should draw customers from, the average of each segment has been calculated on a 5-point scale. The ratings for Segment 1, Segment 2 and Segment 3 are 3.72, 1.52 and 3.00 respectively. It becomes clear that Intel's first priority should be to focus on Segment 1 and Segment 3, while Segment 1 is the most attractive option.

Task 2

Our next step is to create a model that can be used to classify new customers into specific segments. However, not all the variables can be included in the model, because new customers who did not participate in the survey cannot provide importance ratings. The same applies for WTP and income. Additionally, new customers are unlikely to disclose their occupation, media use, or device preference when they fill in personal information required for setting up a new account. Therefore, these variables cannot be used as predictors in the model. Hence, we have

Davide Quaglio 597201
Margarita Tavkazakova 624750
Jinfei Chen 621195
Panagiotis Terzopoulos 629456

decided to initially use five variables for selecting the best model: CompBuy, AmznP, Female, Degree, and Age. These five variables can be easily obtained from the customer's purchasing information. For example, CompBuy can be shown in the account as whether the smartwatch user is a corporate or a private customer. Whether the individual has the Amazon account can also be observable from the personal information. Gender, degree status, and age should be obtained during the account setup process. To build a classification model we will split the raw data into training and test sets (65% - 35% split) and use different methods such as Multinomial Logistic Regression (MLR), Naive Bayes (NB), and Random Forest (RF). The training set will be used to train the models, while the test set will be used to evaluate their performance.

2.1 Multinomial logistic regression

According to the first part of our analysis, our potential customers should be divided into three segments. Multinomial logistic regression can be used as one prediction model to test. We set segment 1 as a reference. The results show that the agreement between the predicted and actual segment membership is 70.57%. From the diagonal elements of the confusion matrix, we observe only 5 values of Segment 1 that were incorrectly predicted as Segment 3 and 8 values of Segment 3 that were incorrectly predicted as Segment 1. However, there is a tendency that Segment 1 and 3 were falsely predicted by Segment 2. The results showed 36 misses for Segment 3 that were incorrectly predicted as Segment 2 and 25 misses for Segment 1 that were incorrectly predicted as Segment 2.

2.2 Naive Bayes

Naive Bayes, a probabilistic machine learning model, can also be used for classification tasks in this case. According to the results, the prediction of Naive Bayes is slightly better than multinomial logistic regression: hit rate of 70.57% vs. 70.86%. As evidenced by the confusion matrix, it is more accurate in predicting Segment 2, i.e. only 6 misses for Segment 2 that were incorrectly predicted as Segment 1 (for MLR, the number is 18).

2.3 Random Forest

Alternative to the previous two methods, the random forest classifier builds an ensemble of models that can classify the data jointly rather than try to fit the single model. The results are as follows: The hit rate reaches 71.14%, which is the highest among the three methods. Also, Random Forest predicted Segment 2 better than the other methods, only 5 wrong predictions as Segment 1 and 7 errors as Segment 3. However, it is highly probable for Segment 1 and 3 to be falsely predicted as Segment 2, 38 and 42 respectively.

2.4 Conclusion

After comparing the accuracy rates of three methods, we can conclude that Random Forest provides the most accurate class predictions, and thus is the most appropriate method for building our classification model. So far, we included five variables in the model: CompBuy, AmznP, Female, Degree and Age. We tried to drop a few variables to see whether the accuracy rates will be affected. After a few trials, the results show that only with three variables i.e. CompBuy, Degree and Age can reach the hit rate of 72.29% with the method of Random Forest. Therefore, we suggest that CompBuy, Degree and Age can be effective predictors.

Furthermore, we could improve the accuracy of prediction by adding more variables as input for the clustering so that a hit rate larger than 72.29% can be reached.

Task 3

Our next aim is to develop a differentiated and sustainable market position and convey our product benefits that are meaningful to the target audience. For that matter, we need to choose partners for launching the new product: Google, Amazon or Aetna. Looking at the

Davide Quaglio 597201
Margarita Tavkazakova 624750
Jinfei Chen 621195
Panagiotis Terzopoulos 629456

characteristics of the first cluster, we can observe that our potential customers care about both health and connectivity to the world (in the form of getting notifications, updates on the traffic, etc.). Google's Android Wear offers a crossover between these two dimensions since it provides the functionality for both wellness and day-to-day tasks. However, a big drawback of this collaboration would be the fact that 69% of the people in the first segment have an iPhone and Android Wear OS works only with IOS 13+ (which has been launched in 2019). Partnership with Amazon would most probably not exploit opportunities coming with high ratings of health-related features in a smartwatch, since its focus lies in optimising task management in daily life. Collaboration with Aetna, on the other hand, will address the need for profound health tracking and the overall trend for paying closer attention to one's physique. We believe that the collaboration of Intel and Aetna will create a valuable product to the companies that are concerned with the well-being of their workers. Intel should position itself to support firms in their endeavours to provide health monitoring for their employees. Why would the companies be interested in offering health smartwatches to their workers? Our targeted audience primarily works in the following industries: finance, tech and sales. According to the research, working in one of these sectors is often associated with poor physical and mental health due to sleep deprivation and exhaustion from working long hours. That in return leads to employee absenteeism which inevitably impacts company's ability to satisfy customer demands and generate revenue. Good health of employees is strategically important to any business performance and Intel smartwatch will help the companies to address this challenge. We will be targeting companies in the above-mentioned industries that are aware of the importance of prioritising the health of their employees and are already offering programs designed to promote fitness for their staff. This way we will be able to reach people who are engaged in sports and won't have to start from scratch trying to explain the significance of exercising. The first step in using our smartwatch will be taking a test that will allow us to gain an understanding of who the user is and what his expectations are. Consequently, the smartwatch will adapt its features depending on the results of the test. We will offer a partnership with fitness facilities to make it easier for people who own our smartwatch to track their workouts and earn rewards for exercising. Each week customers will receive personalised fitness goals (for example, working out a certain amount of time) and will be rewarded regularly if they meet those. We will offer a wide range of rewards starting from discounts on sportwear to free swimming classes for children.

We believe that the most appropriate partner for the second attractive segment we have picked, namely Constant Communicators, is Amazon. More than a half of our respondents use Amazon Prime which indicates their familiarity with the brand and could potentially make them more susceptible to using the result of the collaboration with Amazon. Moreover, the smartwatch will satisfy the need for instant communication and staying up-to-date with current events. The new product will pair well with Bluetooth buds which is very convenient considering the fact that a large majority of the respondents in the selected sector listen to podcasts and radio.

Declaration of Academic Honesty

I, Panagiotis Terzopoulos, hereby declare that I have not previously submitted the present work for other examinations. I wrote this work independently. All sources, including sources from the Internet, that I have reproduced in either an unaltered or modified form (particularly sources for texts, graphs, tables, and images), have been acknowledged by me as such. I understand that violations of these principles will result in proceedings regarding deception or attempted deception.

Ανακτήσιμη υπογραφή

X

Υπογραφή από: b72dfa6c-727d-4fbf-bf39-f1a5920e2857

Panagiotis Terzopoulos

Berlin, February 7, 2023