Group 6: Sinan Wang 620146, Akanksha Saxena 613826, Berivan Teper 618443,
Satyaki Ghosh 613827

Intel is one of the largest manufacturers of computer chips in the world. The company has created a large market for powerful processors. But it's focus on PCs led to ignore the exponential growth of smartphone and smart devices. Recently, Intel decided to enter the smartwatch market and is looking for partner to help it launch and market the product. In this report, we are going to analyse how Intel can effectively segment the market for smartwatches and which segment they should target. Additionally, we also develop a prediction model to classify new customers in segments we got from cluster analysis.

## Checking fit of data for analysis

There are no missing values in the data set that could cause a problem in the analysis. However, there are outliers detected in the data such as the Importance attributes 'ConstCom', 'SaveMT', 'Photo', and 'WTP' which could affect the analysis negatively (Figure 1). Further outliers were seen in the 'Income'. Looking at the importance attributes, one can see that they are more or less normally distributed, which is important for further analysis.

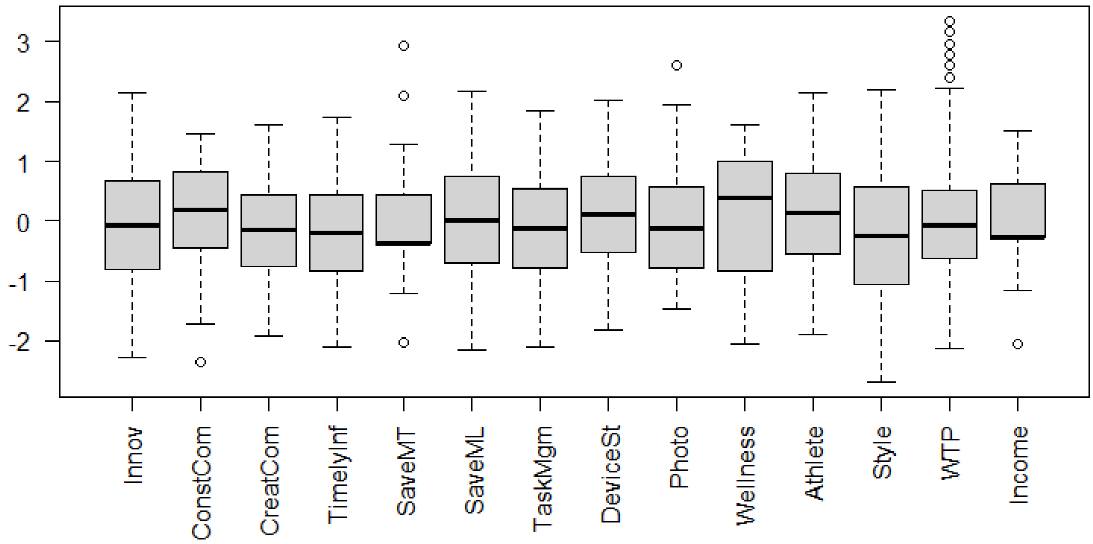


Figure 1: Detected Outliers of Product Attributes, WTP and Income

The distribution of the data can be skewed in the direction of the outlier and make the analysis more difficult. In this case the outliers are considered as contextual outliers, meaning that the outlier is deviating from the rest of the data but within the same context and is not an abnormality. Removing such outliers could alter the results and interpretation, therefore this study is handling outliers by choosing a method that is robust to outliers.

## Describing the data

The total number of respondents is 1000. The participation of female respondents is 56.60% slightly higher than male respondents with 43.60%. The mean age is 35.5, the youngest respondent is 24 and the oldest is 47 years old. 53.40% of all respondents have an iPhone of which 59.00% are female. The remaining respondents (46.60%) most probably own smartphones with Android as their operating system. Slightly above half of all respondents have an Amazon Prime subscription (54.40%). Only 20.10% of the total respondents stated that their companies buy technology devices for them. The largest occupation groups out of these 20.10% are Finance and Sales. The most frequently used media types by the respondents are Facebook and Instagram (74.30%) and TV (79.30%). Around half of all respondents use Twitter, Podcasts and Radios and the Newspaper frequently. Only about 30 percent of all respondents use Snapchat frequently. Occupations among respondents are distributed as follows: Sales (14.00%), Finance (12.80%), Advertising (10.40%), Tech (10.10%), Small Business Owner (9.80%), Education (8.50%), Retail (7.80%), Caretaker (7.50%), Construction (7.40%), Health (6.50%) and Engineering (5.20%). In average respondents show an annual income of $71k to $100k and this is considered to be a middle-class income. The Willingness-To-Pay (WTP) on average is $212.90 for a smartwatch with features they desire.

**Cluster Analysis**

**Deciding on objects and variables**

Our data set consists of 37 variables describing the people's perception of different product features for a smartwatch, and certain characteristics of the population like age, income, occupation, educational qualification, gender, frequently used media channels, and so on. One can argue that the variables can be broadly classified into namely three categories given as follows:

> **Product attributes**: These variables describe the characteristics that the smartwatch by intel can offer like innovation, ability to receive instant messages, getting up-to-the-minute updates of traffic, weather, and so on.
>
> **Willingness to Pay:** The amount of money an individual is willing to pay for the features of the product that are important to him/her.
>
> **Population attributes:** These variables describe the characteristics of the population like age, degree, income, gender, occupation, frequent media channel used, possession of iPhone or other smartphone and whether the company purchases technology for the employee.

For the purpose of clustering, we worked with two set of variables which will act as basis for segmentation, which are as follows:

> **First Set:** Product attributes, Willingness to pay and Population attributes, i.e., we took into consideration all the 37 feature variables in the dataset. It is imperative to investigate all variables for segmentation in cases where one does not have a priori knowledge of the business domain. This can assist in determining which variables are having significant contributions in clustering.
>
> **Second Set:** To understand the needs and preferences of the consumer segments for market analysis one can consider doing cluster analysis on product attributes and willingness to pay. The population attributes will help to describe the groups.

Performing 4-cluster analysis on all variables in the dataset, we realized that there is a high intersection between two of the clusters in key features like all product attributes, willingness to pay, age, income, degree and gender. Social-demographic variables in segmentation seem to not give reasonably distinct clusters.

On the contrary, the 4-cluster solution with product attributes and willingness to pay as predictor variables gives clear and non-overlapping customer segments. Hence, we chose the 'Second Set' of predictor variables for clustering.

**Selecting proximity measure**

The second set of predictor variables consists of product attributes and willingness to pay. The 12 product attributes are measured on a 7-point rating scale and the willingness to pay is a continuous variable with minimum value 100 and maximum value of 390. Taking the assumption that the product attributes are based on interval scale one can select the Manhattan or Euclidean distance as a proximity measure for metric data.

One can standardize the input variables so that they are all on the same scale and more comparable and compute distance-based measures. In the given smartwatch case, the 4-cluster solution is exactly the same for both Euclidean and Manhattan distance in hierarchical clustering. Therefore, it does not make a difference on clustering either of them. One would say that Euclidean distance can be considered as it is the most commonly used distance measure.

**Deciding appropriate algorithm for segmentation**

The single linkage, average linkage, centroid linkage and median linkage methods are sensitive to outliers. As we can see from Figure 1, we have outliers in attributes such as 'ConstCom', 'SaveMT', 'Photo', and 'WTP' in our data. This is causing a chaining effect in the clusters. Hence, we reject them from our analysis. Complete linkage and ward method produced equally sized clusters. Complete linkage can generate distinct clusters, which does not have chaining effect and also has good performance on Variance Ratio Scale for 4 clusters. Similarly, k-means with 4-cluster solution gives a comprehensible segmentation of the population with reasonable size clusters. However, both k-means and complete linkage are highly susceptible to outliers. As an alternative, ward method, which is least susceptible to outliers and noise in the data, seems a sound approach for segmentation in this context.

**Determining the number of clusters**

From the Figure 2, it is shown that the Variance Ratio Criterion (VRC) value decreases with the number of clusters going up. Noticeably, there is a big dip (elbow) in the plot from three to four clusters. Then it continuously decreases but more smoothly. Therefore, this VRC suggests a 3-cluster or 4-cluster solution.

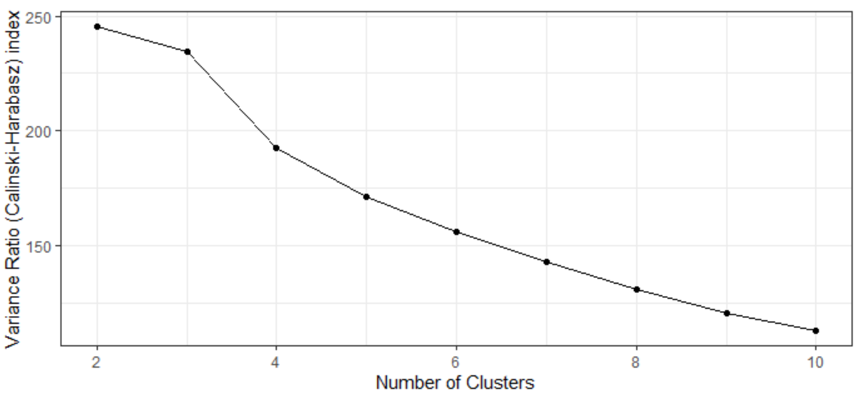


Figure 2: Clustering with Product Attributes and WTP using Euclidean Distance and Ward's Method

In the 3-cluster solution, cluster one is of 40.00% size, cluster two of 19.00% and the third cluster of 41.00%. However, the interpretability of all clusters increases when we aim for a 4-cluster solution. Because the third cluster of the 3-cluster solution is further separated into two clusters of sizes 28.00% and 13.00%. Therefore, one can draw more insightful and distinct interpretations from a 4-cluster solution compared with a 3-cluster solution. Because the line of the VRC is smoothly descending after four clusters, we will take a look at 5- and 6-cluster solutions as well. The segment sizes in the 5-cluster solution are disproportional, one cluster is five times smaller than the others. Moreover, both 5- and 6- cluster solutions have high overlap on predictor variables, age and income of the different groups which hinders distinct clustering. In conclusion, we choose a 4-cluster solution.

**Cluster Descriptions**

We named the 4 clusters as Conventionalists, Workaholics, Trendies, and Fitness enthusiasts.

| | Imp_ Innov | Imp_ ConstCom | Imp_ CreatCom | Imp_TimelyInf | Imp_SaveMT | Imp_SaveML | Imp_TaskMgm | Imp_DeviceSt | Imp_Photo |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 3.76 | 4.05 | 4.21 | 3.57 | 3.49 | 2.91 | 3.19 | 2.67 | 3.66 |
| 2 | 3.97 | 5.35 | 2.35 | 5.62 | 2.08 | 3.82 | 4.97 | 5.25 | 1.92 |
| 3 | 4.73 | 5.32 | 5.37 | 4.49 | 4.10 | 4.74 | 5.54 | 4.05 | 3.23 |
| 4 | 3.91 | 4.33 | 4.76 | 4.11 | 3.93 | 5.84 | 3.17 | 4.91 | 3.30 |
| | Imp_Wellness | Imp_Athlete | Imp_Style | WTP | FB_Insta | Twit | Snap | YouTube | Pod_radio |
| 1 | 3.32 | 3.07 | 3.72 | 190.88 | 0.70 | 0.47 | 0.19 | 0.49 | 0.57 |
| 2 | 3.63 | 2.50 | 3.40 | 251.48 | 0.51 | 0.64 | 0.17 | 0.62 | 0.71 |
| 3 | 5.93 | 4.68 | 5.58 | 230 | 0.89 | 0.38 | 0.54 | 0.51 | 0.61 |
| 4 | 5.24 | 6.12 | 4.57 | 186.43 | 0.88 | 0.58 | 0.81 | 0.85 | 0.40 |
| | TV | NewsP | AmznP | Age | Female | Degree | Income | iPhone | CompBuy |
| 1 | 0.97 | 0.57 | 0.41 | 39.07 | 0.52 | 1.17 | 2.75 | 0.36 | 0.08 |
| 2 | 0.90 | 0.75 | 0.51 | 39.15 | 0.46 | 1.44 | 3.80 | 0.57 | 0.53 |
| 3 | 0.72 | 0.63 | 0.66 | 31.74 | 0.65 | 1.52 | 3.51 | 0.75 | 0.21 |
| 4 | 0.23 | 0.59 | 0.90 | 27.55 | 0.68 | 1.26 | 3.74 | 0.54 | 0.08 |

| | Occup_ Health | Occup_ Finc | Occup_ Sales | Occup_ Advt | Occup_ Edu | Occup_ Cons | Occup_ Eng | Occup _Tech | Occup_ Retail | Occup_ SMB |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.04 | 0.09 | 0.13 | 0.09 | 0.13 | 0.11 | 0.03 | 0.10 | 0.12 | 0.10 |
| 2 | 0.04 | 0.12 | 0.24 | 0.03 | 0.03 | 0.12 | 0.10 | 0.05 | 0.02 | 0.22 |
| 3 | 0.10 | 0.18 | 0.11 | 0.17 | 0.05 | 0.02 | 0.04 | 0.11 | 0.04 | 0.05 |
| 4 | 0.12 | 0.14 | 0.08 | 0.12 | 0.10 | 0.02 | 0.08 | 0.17 | 0.12 | 0.02 |

Table 1: Mean Values Across Each Cluster

Mean values of clusters are shown in Table 1 above and referenced in cluster descriptions.

**Conventionalists**: This is the largest cluster (39.80%). These people were mostly interested in taking photos and having constant and creative communication with their technology devices. They have a low willingness to pay for a smartwatch ($190.88), which goes with their nonchalant attitude towards attributes like style, timely information, device sturdiness etc. The mean age of the group is 39 years, and their income is the lowest ($40k-70k) among all clusters. They are open to new media like Facebook and Instagram (0.70) but mostly dependent on TV (0.97). This cluster has the lowest mean of people holding Amazon Prime subscriptions (0.41) and iPhones (0.36). Based on these characteristics, we decided to call them conventionalists as they are not into innovative and new technologies and do not enjoy changing with time.

**Workaholics**: This cluster corresponds to 18.90 % homogeneity in the population. This cluster has high importance for constant communication, timely information, task management and device sturdiness. Maybe it is because of their occupation. People in this cluster mainly work in Sales (0.24), Transportation (0.12) or are Small/Medium Business Owners (0.22) and have to constantly engage in conversation with other people. They are looking for normal features that would help with their multitasking abilities, receiving instant information on the go (weather, routes) and willing to pay a high amount for it ($251.48). It seems they mainly use technology devices to plan and automate their tasks, to be efficient at work and so it is for this reason we named them workaholics. Moreover, they have the highest income (3.80) which can be approximated to an income range of $100k-175K. They frequently check radio podcast (0.71) and TV (0.90).

**Trendies:** This is the second-largest cluster with a share of 28.40%. People in this cluster stand out with their needs for wellness applications, innovative and stylish technology devices. This cluster has the highest mean value of iPhones (0.75) across all clusters and also has the second-highest Amazon Prime subscriptions (0.66). They have a high income ($71k-100k) and willingness to pay ($230). Their mean age is about 32 years and almost 65% of them are females. Their preferred communication channel is Facebook and Instagram (0.89). They aptly represent the modern generation which uses lots of new and innovative applications in their everyday use of technology devices. This cluster characterizes the young female population who follows trends in technology and fashion. They are thus someone who is following the trend and hence we call them trendies.

**Fitness enthusiasts**: This is the smallest cluster with 12.90% of respondents. They have the lowest age among all clusters with a mean age of 27.55 years and probably that is why they are the group with the highest importance to fitness features in a technology device. Additionally, they also have affinity for wellness features and saving on car, life and health insurance by leading a healthy lifestyle. That is why we named this group fitness enthusiasts. Almost 90% of them have Amazon Prime Subscriptions and 54% have iPhones. 68% of the group is females. Although, this thus seems like an attractive segment to target with a device having suitable features for athletes, their lowest willingness to pay ($186.43) for a smartwatch despite having a moderate-income ($71-100k) causes hindrance in targeting them. The media channel they use is social media like Facebook & Instagram (0.88), Snapchat (0.81), and YouTube (0.85).

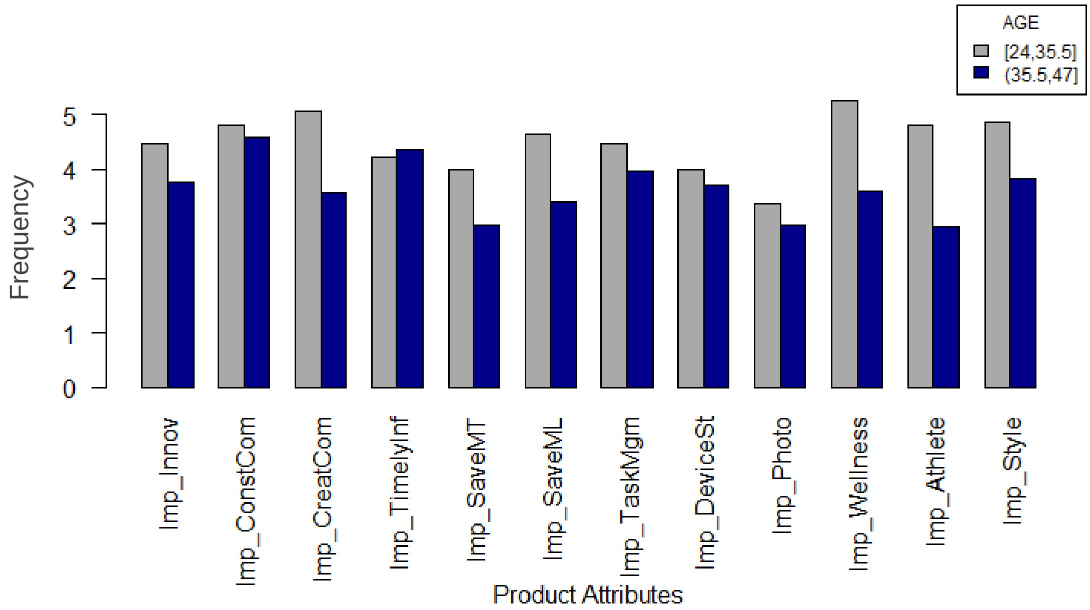


Figure 3: Distribution of Product Attributes and Age

From Figure 3 it can be seen that people in the age group between 35.5-47 have a high preference for product attributes like constant communication, timely information, and task management. This information corresponds to the group of workaholics whose mean age is 39.5. One can see that people between the ages 27-35.5 have a very high preference for wellness, athletics, saving money on ML and stylish smartwatches. This is reflective of groups trendies and fitness enthusiasts. This shows that our clusters are in sync with the overall data trend.

**Market attractiveness of segment**

Considering the attractiveness of the segments, the factors, such as the size of clusters, income, willingness to pay, and importance on innovation play an important role for Intel. The size of clusters has an impact on the sales performance and market share as a whole. Willingness to pay and income are closely related to the ability of the segment to purchase the smartwatch products. Furthermore, Intel has positioned itself as a revolutionist in the technology industry. Therefore, the segment which has a higher affinity towards technological advancement is best suited for the company's market placement.

We plotted a graph between willingness to pay and income using the overall data (Figure 4). One can see that all the income groups are having higher preference to purchase smartwatches in a price range of $175-250. It is also evident that increasing the price of smartwatch over $250 would negatively affect the purchasing behaviour of customers; even those people who have income above $100k are not willing to pay a price higher than $250. People with low income like $40k would also like to pay in a price range of $176-$250. This shows that there is an agreement between all income levels regarding the worth of the smartwatch. People would not like to pay too high or too low for the smartwatch. So, we can say that Intel can market the product within the medium price range $175-250.

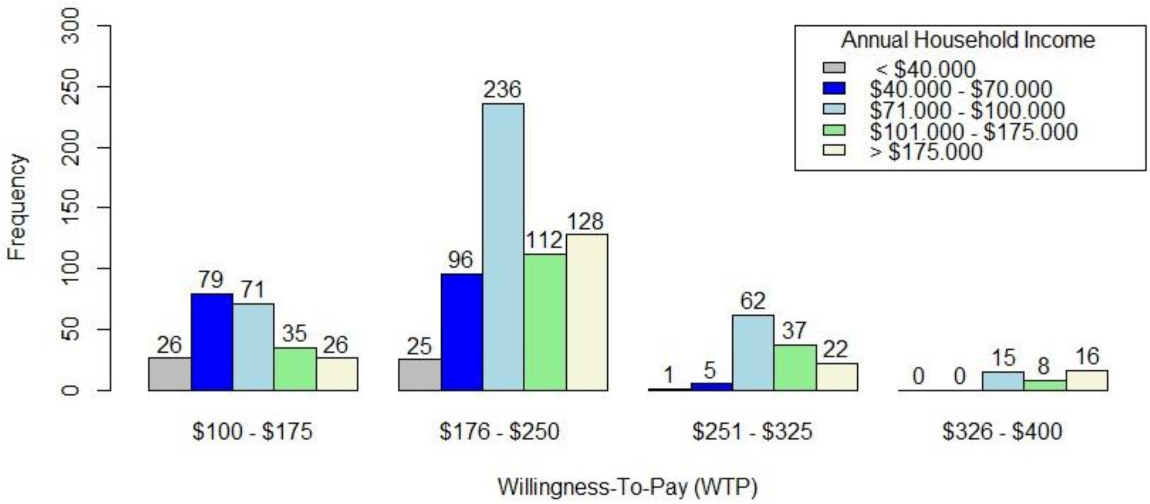


Figure 4: Distribution of WTP and Income

**Conventionalist:** It would have been very beneficial for Intel to target this cluster as it accounts for the largest share of respondents, but it does not correspond well with Intel's competitive advantage.  They seem to be indifferent towards innovative technical features. On a scale of 1-7, 3 is a reasonable score for this cluster. This makes it low on the market attractiveness. Intel could try to engage this segment by advertising features like receiving instant messages at all times without opening phones or taking photos on the run. They could target them through TV advertisements and could also keep the price at a range of $175.

From the Figure 4 one can see that people in the income range $40k-70k, which corresponds to this cluster, would like to pay within the range of $175-250. Therefore, keeping the price at a medium range might be helpful in attracting this segment.

**Workaholics:** This group is high on market advantage and moderate on market attractiveness, making it a good opportunity to further grow the market. Intel can strengthen its competitive advantage in this segment by highlighting product features, which are most important to this group. Intel could better target this segment with TV advertisements. Overall, on a scale of 1-7, 5 is reasonable for this cluster.

**Trendies**: This is the most attractive segment to target among all clusters because they will definitely buy a product with attractive features, and which is stylish and trending in the market. This has high congruence to Intel's vision. This segment is on the lookout for new features and Intel might not want to lose them to competitors. Therefore 6 is reasonable for this segment. To attract this group Intel can partner with Amazon as this collaboration could provide cutting edge technology like a voice-based AI. Furthermore, Amazon also has an obvious advantage in promotion and distribution, as it can utilise its own retail platform. Intel can also engage them with Facebook or Instagram campaigns.

**Fitness enthusiasts:** Although this segment seems like an attractive segment to target because of their high proclivity towards wellness features, their lowest willingness to pay despite a moderate income causes hindrance in targeting them. This is not an attractive segment both in terms of profitability and interest in smartwatch technology. It does not align with Intel's market strength. They might be more towards saving at this point of their life and all their current needs are fulfilled by a smartphone. A smartwatch with more or less the same features as a smartphone might be futile for them. Overall, 2 is reasonable for attractiveness rating.

In order to decide on a smartwatch partner like Aetna, Google or Amazon one needs to look at the potential target segments. According to our attractiveness ratings, we would like to target the trendies and workaholic segments. Both the groups are tech savvy people, one would want a smartwatch for style and innovation, and the other for task management at work. Amazon and Google could be both suitable as partners to target these segments as they both align with Intel's innovation streak. However, both the clusters have above average subscription to Amazon Prime. Also, both these clusters have high mean value on the usage of iPhone. This suggests that people do not prefer working with android or other smartphones. Therefore, it would be more suitable for Intel to partner with Amazon than Google as people in these segments seem to prefer Amazon services.

Based on target segments: trendies and workaholics which comprise almost 50% of population. The features that can be implemented in the smartwatch are communication, task management, wellness applications and the watch can be made stylish. The price for the smartwatch can be set as $230.

## Building a classification model for targeting new customers

### Defining new and potential customers

What has to be admitted is that Intel has a competitive advantage in the computer chip industry and has been established for more than 50 years. There are already many existing customers who have bought digital products from the Intel group. It is logical that Intel has personal information about these customers, and they can also be regarded as new and potential customers for the smartwatch program. Intel can also obtain new customers' information by tracking their activity on the company's website and social media pages or purchase new customer data from an information or data broker legally.

### Selecting predictor variables

As for these new and potential customers, their demographic information, such as gender, age, and education level, is relatively easy to obtain from existing customer databases or data brokers. In contrast, the information about perceptions on importance, willingness to pay is relatively hard to obtain unless Intel conducts a survey. Therefore, the descriptor variables, i.e., iPhone, CompBuy, occupation, media use, AmznP, age, female, degree, income, are served as the main predictors in the classification model. The descriptive analysis which checks outliers and distribution patterns for the predictor variables has been conducted. It shows that they are also suitable to be used as predictor variables.

### Comparing performance of classification methods

Our data consist of four clusters obtained from segmentation analysis done previously. The classification methods used in this report are logistic regression, naive Bayes, and random forest. Since our data contains more than two segments, one can perform a multinomial logistic regression. It uses a logistic function to predict segment membership by taking one segment as a reference category. A Naive Bayes classifier relies on Bayes theorem with strong independence assumptions between the features. A random forest classifier builds an ensemble of decision trees that jointly classify the data.

At first, the data set is split into a training set (60.00%) and test set (40.00%). The training data is used to predict membership as well as possible. The resulting model from training data is then evaluated for performance using the test data. The performance of the classification methods can be compared in terms of hit rate and versus chance. Hit rate defines the agreement between predicted and actual segment membership. Versus chance can be used to test if the estimated accuracy is significantly better than the chance level. After applying these three predict methods the figures are shown in the Table 2 below.

| | Logistic Regression | Naive Bayes | Random Forest |
|---|---|---|---|
| Hit rate | 77.25% | 70.50% | 78.50% |
| Versus chance (adjust for base rate) | 49.71% | 36.39% | 52.20% |

Table 2: Comparison of three classification methods

In logistic regression, it is shown 77.25% on hit rate and performs equally with versus chance adjusted for base rate. As for the naive bayes method, one can argue that the hit rate is 70.50% and performs worse than the chance adjusted for base rate. Compared with logistic regression and naive bayes, random forest has the biggest hit rate with 78.50%, which is a relatively accurate result for prediction. And random forest performs 2.20% better than versus chance adjusted for base rate. Overall, in this context random forest performs the best as for prediction.

**Interpretation of selected random forest**

Random forest model performs the best in this context. Random forest splits the data into training and 'out of bag' (OBB) and optimises each tree on its own training data. The OBB estimate of error rate is 16.50% in this case. And the confusion matrix of random forest's own training data is shown in the Table 3 below.

| | Conventionalists | Workaholics | Trendies | Fitness enthusiasts | Class. error |
|---|---|---|---|---|---|
| Conventionalists | 236 | 10 | 7 | 0 | 0.07 |
| Workaholics | 30 | 70 | 9 | 0 | 0.36 |
| Trendies | 24 | 3 | 139 | 3 | 0.18 |
| Fitness enthusiasts | 2 | 0 | 11 | 56 | 0.19 |

Table 3: Confusion matrix of random forest

As suggested in the table, the model has the lowest error rate (7.00%) in classifying conventionalists, which means that 236 out of 253 memberships are accurately classified. However, the model has relatively the highest error (35.00%) on classifying workaholics and a similar error rate, around 20.00%, in classifying Trendies and Fitness enthusiasts.