# Project 1

## Task 1

Variable ‘respondent\_id’ is an identifier, and has a unique value for each respondent.

‘variety\_of\_choice’ is an attitudinal variable that indicates the respondent’s opinion on the variety of choice when shopping for office equipment. The variable has a minimum value of 4, a maximum value of 10, with a mean of 7.565. This indicates that this group of respondents value variety of choice, as this attitudinal variable has the highest variable mean.

‘electronics’ is an attitudinal variable that indicates the respondent’s opinion on electronics availability when shopping for office equipment. The variable has a minimum value of 1, a maximum value of 10, with a mean of 4.45.

‘furniture’ is an attitudinal variable that indicates the respondent’s opinion on furniture availability when shopping for office equipment. The variable has a minimum value of 0, a maximum value of 9, with a mean of 3.27. This indicates that this group of respondents do not value furniture availability, as this attitudinal variable has the lowest variable mean.

‘quality\_of\_service’ is an attitudinal variable that indicates the respondent’s opinion on service quality when shopping for office equipment. The variable has a minimum value of 1, a maximum value of 9, with a mean of 3.53. This variable has a 3rd quartile value of 4, indicating that a small group of customers find service quality very important.

‘low\_prices’ is an attitudinal variable that indicates the respondent’s opinion on prices when shopping for office equipment. The variable has a minimum value of 1, a maximum of 10 and a mean of 4.795.

‘return\_policy’ is an attitudinal variable that indicates the respondent’s opinion on the presence of a return policy when shopping for office equipment. The variable has a minimum of 1, a maximum of 10, with a mean of 4.25.

‘professional’ variable describes whether the respondent is a professional or not. Only 68 out of 200 respondents are professionals (34%).

‘income’ variable indicates how much the respondent earns in thousands per year, in pound sterling. While the mean for ‘income’ is 32.17, the median is 19.5, indicating that most respondents earn very little, and a few respondents earn substantially more, with a maximum of 95.

‘age’ variable indicates the respondent’s age. While the mean for ‘age’ is 32.52, the median is 27, indicating that respondents are concentrated below the age of 30 with the other half of respondents spread between the ages of 30 and 68.

## Task 2

I created a new dataframe ‘data\_att’, where I isolated the attitudinal variables of the original dataset. I considered attitudinal variables to be the six variables ‘variety\_of\_choice’, ‘electronics’, ‘furniture’, ‘quality\_of\_service’, ‘low\_prices’, and ‘return\_policy’.

I normalised these variables to create a new dataframe, ‘data\_att\_norm’, to contain the normalised versions of these values. I used z-score standardisation as the method of normalisation for all variables.

The ‘electronics’ variable had both the smallest minimum value and the largest maximum value across all six normalized variables, at -1.77534 and 2.85598 respectively. Other minimums and maximums can be observed in Figure X.

## Task 3

To perform hierarchical clustering on the normalised attitudinal variables, I first calculated the Euclidean distances between observations.

Chart

Description automatically generated with medium confidenceUsing these distances, I then performed hierarchical clustering with the appropriate library and the method “ward.D2”. The dendogram of the algorithm’s resulting clusters can be seen in Figure X.

Figure X: Dendogram for hierarchical clustering algorithm.

## Task 4

Observation numbers for each cluster of a six-cluster solution can be seen in Table X. 59 observations were assigned to the largest cluster, 1, and only 8 observations were assigned to the smallest cluster, 2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 |
| 59 | 8 | 52 | 17 | 35 | 29 |

Table X: Number of observations assigned to each cluster for a six-cluster solution.

## Task 5

Chart, bar chart

Description automatically generatedThe segment profile plot, for a six-cluster solution, generated by the ‘flexclust’ package can be observed in Figure X.

Figure X: Segment profile plots for each cluster in a six-cluster solution.

The table of cluster memberships can be observed in Table X. The results indicate that ‘hclust’ and ‘as.kcca’ procedures are almost in full agreement, as only 2 observation clusters were reassigned as a result of the procedures. The reassigned observations can be found in Table X.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 0 | 59 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 8 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 52 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 17 |
| 5 | 33 | 2 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 29 | 0 |

Table X: Cluster assignments for ‘hclust’ (vertical) and ‘as.kcca’ (horizontal) procedures.

## Task 6

Cluster 1 was assigned 59 observations in total. Customers of this cluster value variety of choice and furniture more than the average customer, but do not care for return policies, low prices, or quality of service as much as the average customer.

Cluster 2 was assigned 8 observations in total. Customers of this cluster value low prices a lot more than the average customer, and they did not find it important whether an office supply store had furniture, compared to the average customer. They also did not find return policies important.

Cluster 3 was assigned 52 observations in total. Customers of this cluster value return policies and low prices significantly more than average. These customers, when compared to the average, cared less about all other aspects of the store.

Cluster 4 was assigned 17 observations. These customers find it important that their office supplies store carries electronics, and they find return policies to be even more important. However, customers of this cluster do not care for variety of choice and are not looking for low prices.

Cluster 5 was assigned 35 observations. These customers find electronics, variety of choice and furniture to be important aspects of an office supplies store. They do not care as much about quality of service and found return policies to be less important when compared to the average.

Cluster 6 was assigned 29 observations. Customers of this cluster found everything to be of little importance when compared to quality of service. This cluster’s customers rated quality of service importance to be the highest out of any other cluster.

## Task 7

Cluster 2 only contains eight observations, which makes it a significantly smaller cluster when compared to all other clusters. This indicates that the cluster membership is very specific, which are less useful when observing customer behaviour as we are attempting to generate a generalised solution. Hence, it would be wiser to observe the results of a five or less cluster solution, where the cluster’s observations could be absorbed by other clusters.

## Task 8

Observation numbers for each cluster of a five-cluster solution can be seen in Table X. 60 observations were assigned to the largest cluster, 2, and 17 observations were assigned to the smallest cluster, 3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
| 59 | 60 | 17 | 35 | 29 |

Table X: Number of observations assigned to each cluster for a five-cluster solution.

## Task 9

Chart, bar chart

Description automatically generatedThe segment profile plot for a five-cluster solution, generated by the ‘flexclust’ package, can be observed in Figure X.

Figure X: Segment profile plots for each cluster in a five-cluster solution.

Cluster 1 was assigned 59 observations. I have chosen to name this cluster ‘Interior Designers’, as they value variety of choice and furniture in an office supply store, indicating that they enjoy decorating with a wealth of choice.

Cluster 2 was assigned 60 observations. I have chosen to name this cluster ‘Serial Shoppers’ as they value low prices and return policies, indicating that they buy and return products frequently.

Cluster 3 was assigned 17 observations. I have chosen to name this cluster ‘Tech Reviewers’, as they value electronics and return policies, indicating that they enjoy trying electronics and returning them

Cluster 4 was assigned 35 observations. I have chosen to name this cluster ‘Amazon Shoppers’, as they value electronics, variety of choice and furniture, indicating that they enjoy shopping various categories of products but do not care for quality of service.

Cluster 5 was assigned 29 observations. I have chosen to name this cluster ‘Luxury Seekers’, as they value quality of service significantly and do not care for other aspects of the store, indicating that they are happy with their store experience if they are treated well by the store.

## Task 10

The five-cluster solution is better because the smallest cluster in the six-cluster solution was absorbed by another larger cluster, providing for more generalised insights over the data.

## Task 11

## Task 12

Observation numbers for each k-means cluster of a five-cluster solution can be seen in Table X. 61 observations were assigned to the largest cluster, 5, and 17 observations were assigned to the smallest cluster, 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
| 60 | 17 | 33 | 29 | 61 |

Table X: Number of observations assigned to each cluster for a k-means five-cluster solution.

## Task 13

The hit rate between the two procedures is 99%. Only two observations were assigned different clusters, which is a behaviour that was also apparent in ‘as.kcca’ method, where two observations were assigned to a cluster different than the hierarchical cluster assignments.

# Project 2

## Task 1

## Task 2

The coefficient estimate of ‘discountyes’ for model ‘m1’ is 1.10090. The sign is positive. The coefficient is statistically significant to 10^-3 level.

## Task 3

The odds ratio for ‘discountyes’ is 3.006872. This means that when a user of the ecommerce website was offered a 10% discount on their first order, they were 3 times more likely to make a purchase, compared to users that were not offered this discount.

## Task 4

The 95% confidence interval for the odds ratio for ‘discountyes’ is between 2.8071755 and 3.2205543. This means that the middle 95% of users were 2.8 to 3.2 times more likely to make a purchase if they were offered a discount, compared to users that were not offered this discount.

## Task 5

Chart, bar chart

Description automatically generatedThe double decker mosaic plot featuring the impact of discount on conversion, for each ‘source’ type, can be observed in Figure X.

Figure X: Double decker mosaic plot for effect of discount over conversion, by source type.

Discounts appear to have the most effect on conversion when the user source type is ‘search’. Assuming type ‘search’ means that the user arrived on the ecommerce site through a search engine query, this finding can be used to target users who arrive on the ecommerce site through this channel, and to offer discounts specifically for them.

For user source type ‘ads’, the discount offer appears to have had little impact at best, indicating the store users do not care for discounts when they arrive on the ecommerce site through an advertisement.

Finally, for user source type ‘direct’, the discount offer appears to have had some impact, but less than the impact discounts had on ‘search’.

## Task 6

The coefficient estimate of ‘sourcedirect’ for model ‘m2’ is 0.69434. The sign is positive. The coefficient is statistically significant at 99.999% level.

The coefficient estimate of ‘sourcesearch’ for model ‘m2’ is 0.73645. The sign is positive. The coefficient is statistically significant at 99.999% level.

## Task 7

The odds ratio for ‘sourcedirect’ is 2.00238016. This means that if the ecommerce user accessed the ecommerce website directly, they were 2 times more likely to make a purchase, compared to users that accessed the ecommerce website through another source.

The odds ratio for ‘sourcesearch’ is 2.08849965. This means that if the ecommerce user accessed the ecommerce website through a search, they were 2 times more likely to make a purchase, compared to users that accessed the ecommerce website through another source.

## Task 8

The coefficient estimate of ‘discountyes:sourcedirect’ for model ‘m3’ is 0.30445. The sign is positive. The coefficient is statistically significant at 90% level.

The coefficient estimate of ‘discountyes:sourcesearch’ for model ‘m3’ is 1.30923. The sign is positive. The coefficient is statistically significant at 99.999% level.

## Task 9

The 95% confidence interval for the odds ratio for ‘discountyes:sourcedirect’ is between 0.9561778 and 1.9205866. This means that the middle 95% of users were 0.95 to 1.92 times more likely to make a purchase if they were offered a discount and if they accessed the ecommerce website directly, compared to users that were not offered this discount and/or accessed the ecommerce website through other channels.

The 95% confidence interval for the odds ratio for ‘discountyes:sourcesearch’ is between 2.7653813 and 4.9569491. This means that the middle 95% of users were 2.76 to 4.95 times more likely to make a purchase if they were offered a discount and if they accessed the ecommerce website through a search, compared to users that were not offered this discount and/or accessed the ecommerce website through other channels.

## Task 10

The variables ‘visit\_duration’, ‘sourcedirect’, and ‘discountyes:sourcedirect’ were the only coefficients significant at a 95% level.

## Task 11

The correlation between ‘total\_pages\_visited’ and ‘visit\_duration’ is 0.9993129, which means that the two variables are almost entirely positively correlated. These two models are understandably correlated; logically if a user visits more pages, their visit duration will also be longer.

However, this may result in multicollinearity in our model, where the independent variables of our model are not independent from each other. Multicollinearity can impact the accuracy of coefficient estimates.

## Task 12

For model ‘m5’, ‘total\_pages\_visited’ has a coefficient estimate of 0.416732. This coefficient estimate is also statistically significant at a 99.999% level.

For comparison, ‘total\_pages\_visited’ had a coefficient estimate of -0.054773 in model ‘m4’, which was not statistically significant.

The changes made in model ‘m5’ can overall be deemed positive, as the impact of ‘total\_pages\_visited’ is modelled more accurately.

## Task 13

Chart, box and whisker chart

Description automatically generatedThe visualisation of odds ratios and confidence intervals for the independent variables of model ‘m5’ can be observed in Figure X.

Figure X: Odds ratios (points) and confidence intervals (error bars) of each variable of model ‘m5’.

## Task 14

The mean value of ‘base\_probs’ is 0.1759163.

## Task 15

Using a threshold value of 0.5 for the indicator variable ‘pred\_conversion’, 1358 users were predicted to convert.

## Task 16

‘pred\_conversion’ has an accuracy of 84.237% over ‘conversion’.

## Task 17

AUC metric for predictions made with model ‘m5’ is 0.774793.

## Task 18

Upon increasing all users’ ‘total\_pages\_visited’ metric by one unit, i.e., one more page visited by all users, and adding the new probabilities generated by the model to the dataset, the mean of the probabilities is 0.1759163, which is identical to the figure calculated in Task 14.

## Task 19

Lift calculation for the difference between model ‘m5’ and the hypothetical scenario where all users visited an additional page yields the value 2.00377e-13, as probability calculations for the two models do not change much. The value is very close to zero because a change that impacts all values of a variable equivalently should not change the probability calculations from the previous model.

# Project 3

## Task 1

## Task 2

A relevel is required for the variables ‘cloud\_storage’ and ‘price’. This is because when dummy variables are added to variables of type factor such as the aforementioned variables, the dataset splits factors into dummy variables for each level except for the base level. Defining a base level allows the model to display coefficients for levels other than the base level, thus making coefficients, or in this case the upgrades to features, easier to interpret.

## Task 3

The mean of the newly created ‘price\_n’ variable is 12.01533.

## Task 4

Respondents within the dataset chose the ‘30gb’ option 830 times out of the 3000 total choices.

20.8% of the respondents’ choices had ‘email’ for their ‘cloud\_services’ variable.

## Task 5

Dataset ‘m\_data’ contains 8 columns.

## Task 6

The coefficient estimate of ‘cloud\_storage5000gb’ for model ‘model1’ is 0.894883. The estimate is statistically significant at a 99.999% level.

The coefficient estimate of ‘pricep12’ for model ‘model1’ is -0.836795. The estimate is statistically significant at a 99.999% level.

## Task 7

The coefficient estimate of ‘price\_n’ for model ‘model2’ is -0.133936. The estimate is statistically significant at a 99.999% level. Since this coefficient estimate is continuous, the coefficient estimate indicates that if all other variables stayed constant, and price was increased by one unit, the target variable would decrease by 0.133936. This is different from the ‘price’ variable coefficients in model ‘model1’, as those coefficients could only measure the effect of ‘price = p12’ and ‘price = p18’ on the target variable.

## Task 8

The models ‘model1’ and ‘model2’ are not very different from each other, as their log likelihood is only apart by 0.2. However, model ‘model1’ has the slight edge over model ‘model2’ in log likelihood, and thus is the better choice.

## Task 9

Upon predicting the choice probabilities of different alternatives, the predicted probability of choosing the third alternative in the first choice set is 0.0284, or 2.84%.

## Task 10

Upon computing the predicted alternatives for each choice using the maximum choice probabilities, the predicted alternative in the third choice set is 2.

## Task 11

Upon extracting the selected alternatives for each choice set, the selected alternative in the fifteenth choice set is 2.

## Task 12

The confusion matrix for model ‘model2’ can be observed in Table X.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| 1 | 579 | 211 | 197 |
| 2 | 190 | 624 | 200 |
| 3 | 185 | 200 | 614 |

Table X: Confusion matrix for model ‘model2’.

The accuracy of model ‘model2’ is 0.6056667, or 60.57%. Compared to the expected baseline random prediction accuracy of 33%, model ‘model2’ is almost twice as good at making correct predictions.

## Task 13

Coding task.

## Task 14

Coding task.

## Task 15

The predicted market share for alternative four of the hypothetical market is 0.1445737, or 14.46%.

## Task 16

Upon modifying the ‘cloud\_services’ attribute for the fifth alternative in the hypothetical market, the predicted market share for alternative four increased to 0.1867029, or 18.67%.

## Task 17

Changes in market share between the first hypothetical scenario and the new hypothetical scenario can be observed in Table X.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Alternative 1 | Alternative 2 | Alternative 3 | Alternative 4 | Alternative 5 |
| 0.05245839 | 0.04419348 | 0.03846904 | 0.04212921 | -0.17725012 |

Table X: Changes in market share for each alternative, from first hypothetical scenario to the new hypothetical scenario.

Alternative 5’s market share changed the most, as its market share decreased 17.72%. The change made in the new hypothetical scenario was essentially a downgrade for alternative 5 in terms of features without a change in price, and thus the market share figures accurately reflect the change as alternative 5 becomes less desirable for the price.

## Task 18

Using the model ‘model2’ coefficients, a consumer would be willing to pay £3.68 more per month for customer support.

## Task 19

Using the model ‘model2’ coefficients, a consumer would be willing to pay £1.23 more per month for an upgrade from 30GB to 2000GB cloud storage.

## Task 20

Using the model ‘model2’ coefficients, a consumer would be willing to pay £ 5.45 more per month for an upgrade from 2000GB to 5000GB cloud storage.