Script

Slide 1

1. This dataset is from UCI.
2. In further data inspection, this dataset has 165474 instances and 14 attributes which shows the various variable in buying from the e-commerce.
3. These are the variables collected such as the country they are from, clothes they bought and etc.
4. Our target variable is “price.2” with levels 1 and 2
5. Moreover, “price” is a continuous variable and we normalised it in some of our models to improve the accuracy
6. All the variables will act as predictors for “price.2” which represents “1” as bought higher than average prices and “2” as bought at lower than average prices.
7. We want to know how important each predictors are in predicting items bought at higher or lower than average price

Slide 2

1. This table shows the explanation of each variable

Slide 3

1. We also plotted the histograms for each variable to visualise our data
2. So most data were collected from Poland
3. Our target variable “price.2” has ratio of 84695:80779
4. From the graphs, we can also see that “page.2..clothing.model.” & “colour” skewed to the left

Slide 5

1. Next, we study the distribution of each column by studying the skewness and kurtosis
2. A normal distribution will have skewness of 0 and kurtosis of 3
3. “country” variable is skewed to the left, whereas “Model.photography” & “page” are skewed to the right
4. The target variable “price.2” is almost normally distributed
5. In terms of kurtosis, kurtosis of >3 have longer & fatter tails, and vice versa
6. “Country” & “page” have longer and fatter tails, whereas The rest have thinner and shorter tails.

Slide 6

1. Afterwards, we study the correlation of each column pairs.
2. As a rule of thumb, +ve reflect DIRECT Relation, whereas -ve value reflect INVERSE Relation
3. Correlation value of >0.50 value indicates correlation
4. After tabulating the column pairs and their correlation, “Page.1..main.category.” & “page.2..clothing.model.” has a Strong Direct Correlation, whereas “Price” & “price.2” has a Strong Inverse Correlation

Slide 7

1. Afterwards, we study the correlation of each column pairs.
2. As a rule of thumb, +ve reflect DIRECT Relation, whereas -ve value reflect INVERSE Relation
3. Correlation value of >0.50 value indicates correlation
4. After tabulating the column pairs and their correlation, “Page.1..main.category.” & “page.2..clothing.model.” has a Strong Direct Correlation, whereas “Price” & “price.2” has a Strong Inverse Correlation

**Slides 9: KNN**

1. KNN is non-parametric method which uses Euclidean Distance for classification and regression
2. We used k-fold validation to train the dataset. K value
3. A smaller K value will result in noises to influence the prediction, whereas a higher K-value might result in underfitting the data
4. To choose the optimal k value, we use the square root method

**Slides 11: KNN**

1. Since using the normal KNN method gives us too many ties, we used the weighted KNN method together with the K-fold validation method to train build the model.
2. K value selected was 353 after square-rooting the number of instances in training dataset, and applied a 10-fold validation method to build the model
3. Hence, the overall accuracy obtained was 99.35%

**Slides 12: HC**

1. Afterwards, we performed hierarchical clustering which creates clusters by arranging and classifying the data according to the inclusiveness of hierarchy
2. We used agglomerative clustering which uses the Euclidean and the complete linkage to build a multilevel hierarchy from points to clusters.
3. From what we indicated by the highest average silhouette, we have identified that k=2 is the desirable number.
4. Thus, we further add a border around the two largest clusters which k=2 that is shown in the dendogram below

**Slides 13: HC**

1. Classify Group 1 as “1” & Group 2 as “2” as levels of “price.2”
2. By applying the hierarchical clustering method, we obtained an accuracy of 50.267%.