1. Objective

The overall objective is finding the segments that exist among online shoppers using the data found at “https://archive.ics.uci.edu/ml/machine-learning-databases/00468/”. Furthermore, is the prediction of customers that will buy from the online store.

2. Exploration and Data Analysis

2.1. Data Description

The data has 18 attributes in which 8 are categorical and ten are numerical. The Columns are given in Table 1. A detailed attribute information can be found at “https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset”.

From the file of the data, the column names are 'Administrative', 'Administrative_Duration', 'Informational', 'Informational_Duration', 'ProductRelated', 'ProductRelated_Duration','BounceRates', 'ExitRates', 'PageValues', 'SpecialDay', 'Month', 'OperatingSystems', 'Browser', 'Region', 'TrafficType', 'VisitorType','Weekend' and 'Revenue'. Table 1 shows the columns and corresponding data description.

Table 1: Column Name and Data Values of Data Set

Column Name	Data
Administrative	Continuous - number of administrative pages visited by the visitor
Administrative_Duration'	Continuous - total time spent on administrative page
Informational	Continuous - number of informational pages visited by the visitor
'Informational_Duration'	Continuous - total time spent on informational page
ProductRelated	continuous - number of product pages visited by the visitor
'ProductRelated_Duration'	continuous - total time spent on product page
BounceRates	percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session
ExitRates	The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session
PageValues	the average value for a web page that a user visited before completing an e-commerce transaction
SpecialDay	the closeness of the site visiting time to a specific special day (e.g., Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction
Month	Continuous – months of the year (January through December)
OperatingSystems	continuous
Browser	continuous
Region	continuous
TrafficType	continuous
VisitorType	returning visitor, new visitor, other
Weekend	True, False
Revenue	True, False

The class label (y) will be “Revenue”. It is chosen as the value to be predicted which will indicate if revenue was made from a visitor to the site.

The number of values in each column is shown in Table 2. It shows that the amount of data in each colu mn is 12329 and there are no missing values.

Table 2: Information of the Data Set

```
#   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  object
 11  OperatingSystems         12330 non-null  int64
 12  Browser                  12330 non-null  int64
 13  Region                   12330 non-null  int64
 14  TrafficType              12330 non-null  int64
 15  VisitorType              12330 non-null  object
 16  Weekend                  12330 non-null  bool
 17  Revenue                  12330 non-null  bool
dtypes: bool(2), float64(7), int64(7), object(2)
```

2.2. Visualizations

For exploratory data analysis, some visualizations are done. Figures 1 shows the visitor-type variable. It can be seen that more revenue is made from returning visitors. Figure 2 shows that higher revenues were made in September, October and November.

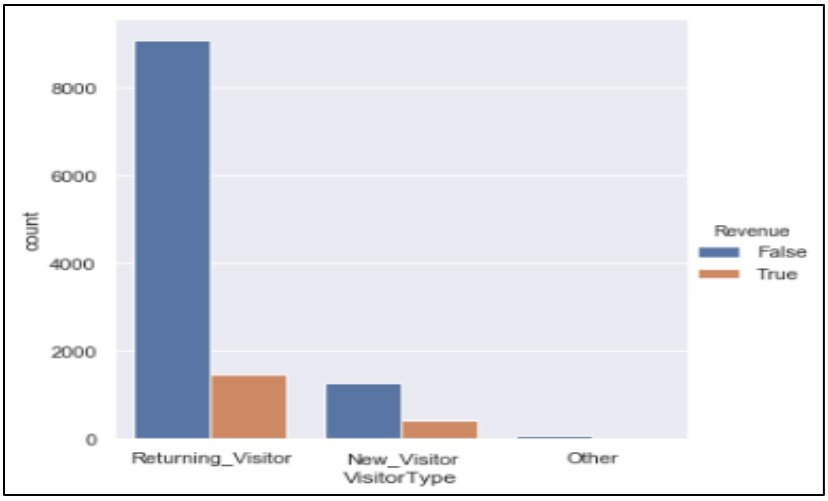

Figure 1: Visitor Types

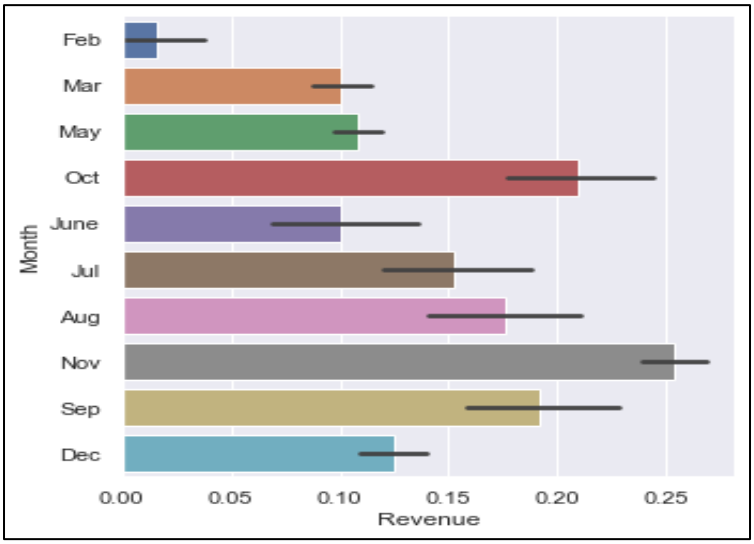

Figure 2: Revenue by months

In order to determine the segments of people that exist withing those from whom revenue are made, the data is grouped by the revenue variable and those with “Revenue =1” is separed for further visualization. From the positive revenue data, figure 3 and 4 shows that a lot of returing visitors sppent time on product related and administrative pages. Figure 5 shows that low revenue was made on weekends. Figure 6 butresses the earlier discovered pattern that higher revenues were made in November, December and May- this can suggest that more revenues were made during the shopping season and closer to holidays.

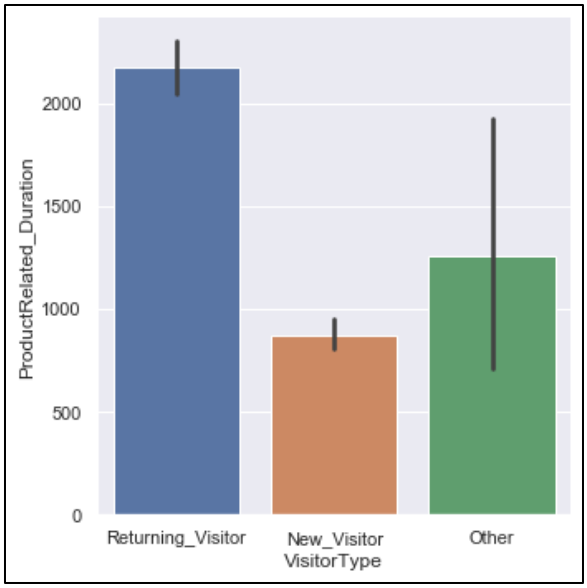

Figure 3: Time spent by visitors on Productrelated sites

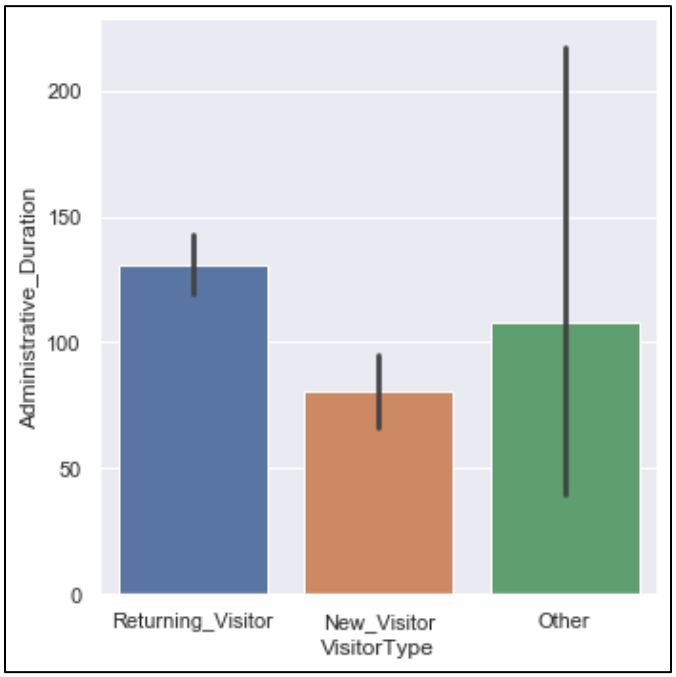

Figure 4: Time spent by visitors on administrative sites

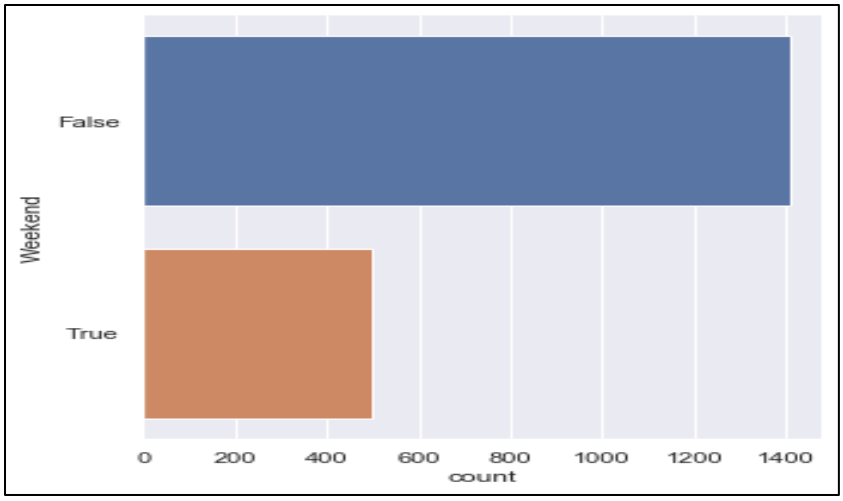

Figure 5: Revenue generated on weekend.

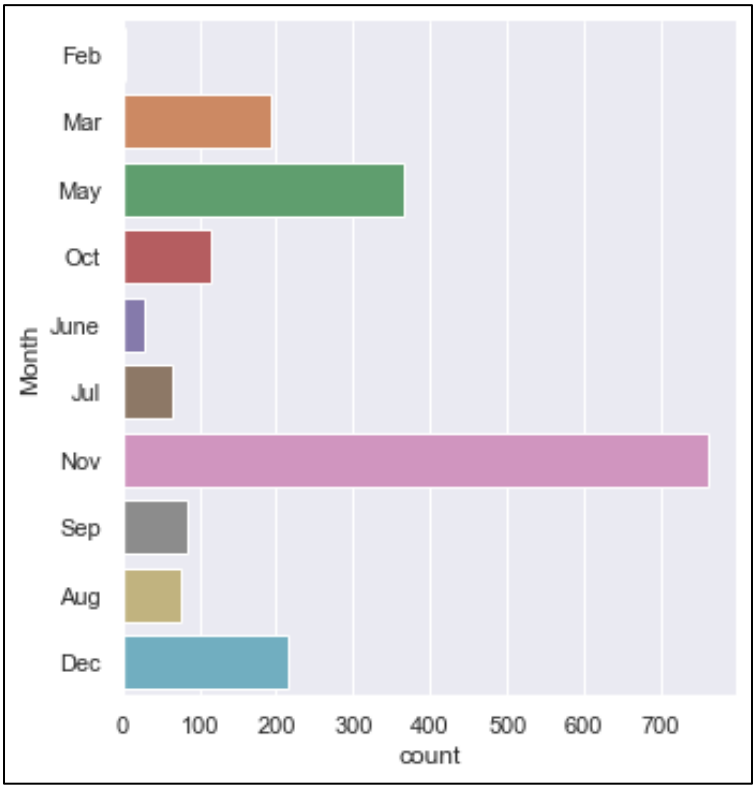

Figure 6: Revenue made in each month

3. Feature Engineering

While working with the segmented data, the feature engineering will focus on converting the non-categorical values to numerical ones. The “Weekend” and “Month” columns converted to numerical columns using label encoder while the ‘one-hot-encoder’ using ‘get dummies’ method is used for the “VisitorType columns. After Feature engineering, the data info obtained is shown in Table

Table 3: View of Data Info after Feature Engineering

```
Data columns (total 19 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Administrative                  1908 non-null   int64
 1   Administrative_Duration         1908 non-null   float64
 2   Informational                   1908 non-null   int64
 3   Informational_Duration          1908 non-null   float64
 4   ProductRelated                  1908 non-null   int64
 5   ProductRelated_Duration         1908 non-null   float64
 6   BounceRates                     1908 non-null   float64
 7   ExitRates                       1908 non-null   float64
 8   PageValues                      1908 non-null   float64
 9   SpecialDay                      1908 non-null   float64
 10  Month                           1908 non-null   int32
 11  OperatingSystems                1908 non-null   int64
 12  Browser                         1908 non-null   int64
 13  Region                          1908 non-null   int64
 14  TrafficType                     1908 non-null   int64
 15  Weekend                         1908 non-null   int64
 16  VisitorType_New_Visitor         1908 non-null   uint8
 17  VisitorType_Other               1908 non-null   uint8
 18  VisitorType_Returning_Visitor   1908 non-null   uint8
dtypes: float64(7), int32(1), int64(8), uint8(3)
```

4. Pre-Modelling

The data is split into training and testing set without having a dependent variable (y) and independent variable (X) since the data being currently worked on is that of only positive revenue section as shown in Table 4 with 30% allocated to the test set.

Furthermore, the training data is scaled

Table 4: Number of rows in Training and Test Sets

Train	Test
1335	573

Furthermore, the train data is scaled.

5. Modelling

K-means is done in order to know the kind of cluster that exists within customers that made purchase. Plotting K-means inertial against the number of clusters checked, figure 7 is obtained. The number of clusters chosen is 4.

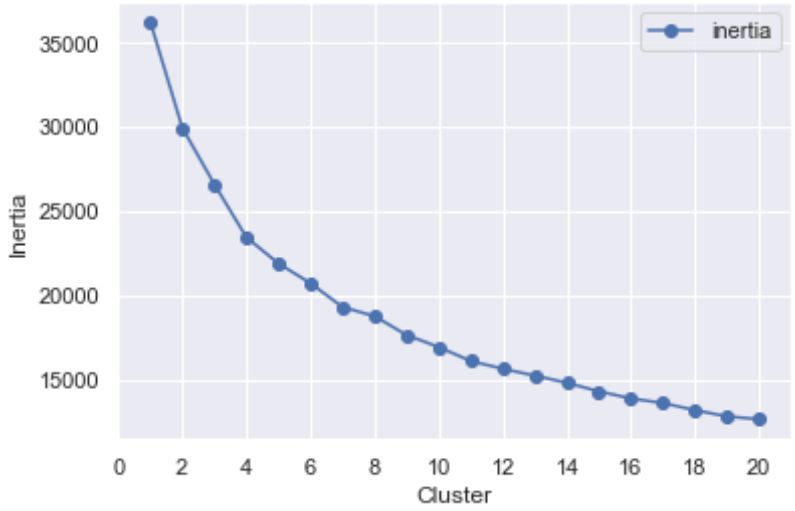

Figure 7: K-mean inertial versus number of clusters

Thus, the K-means model with 4 number of clusters is fitted to the positive revenue data. The model labels (Km.labels_) is extracted and concatenated with the positive revenue data. This concatenated data is grouped by the K-means labels and mean values of each attribute, resulting in 4 segments.

Look through the means the attributes of each section leads to four groups named 'returnees' (returning visitors), 'special day visitors' (other visitors with high traffic on special days and weekends), "year end" (new visitors that come toward the end of the year) and 'product focused' (returning visitors with high production duration).

Principal Component Analysis is the second model which is done to reduce to dimensions of the revenue positive data. The number of components (n), individual explained variance of each component (ind_var) and the cumulative explained variance (var) is plotted in figure 8 and 9. From the figures, the number of components for the PCA model is chosen to be 5 (which explains 59% variance).

Thus, the PCA model using the positive revenue data is made up of 5 components.

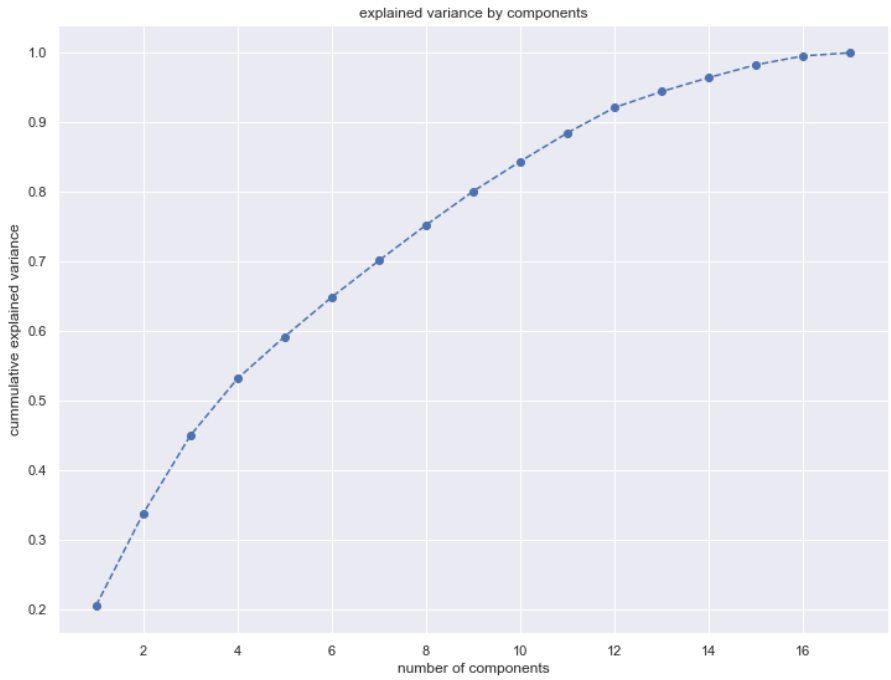

Figure 8: cumulative explained variance against number of components

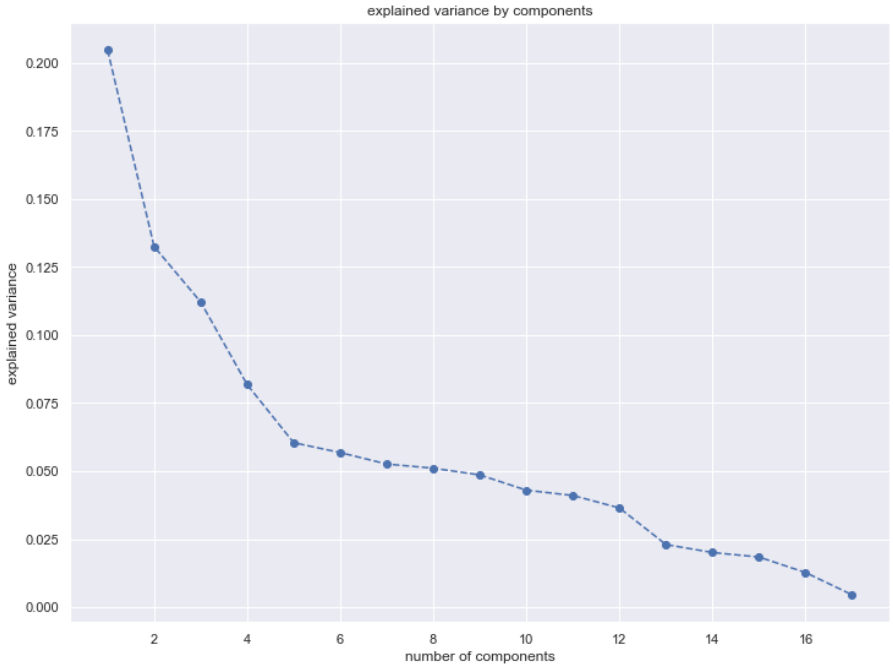

Figure 9: Explained variance of each component

Having built a model with just 5 components instead of 17 attributes, the k-means is run on the reduce model i.e., the K-means is run on the PCA model. The inertial versus cluster figure for the K-means on PCA model is displaced in Figure 10.

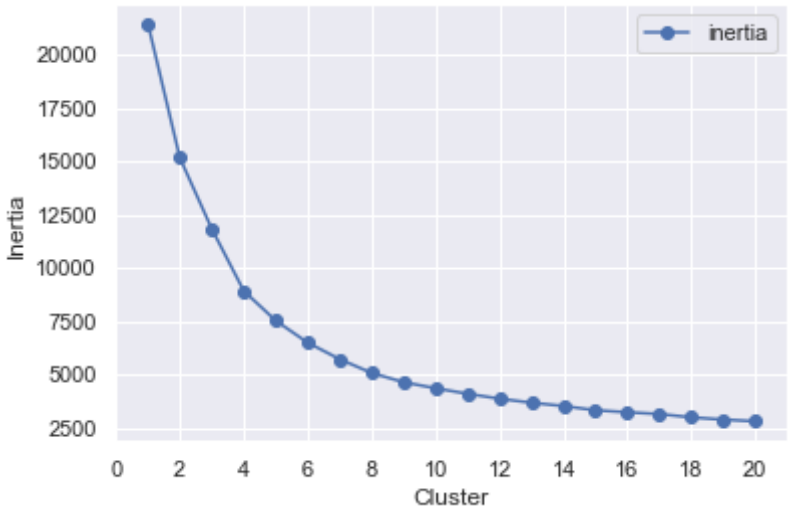

Figure 10: K-mean inertial versus number of clusters for the k-means on PCA model

As before, the number of clusters is chosen as 4. The data is grouped as described above in the k-means section. Thus, four sections arise.

Since 4 dimensions cannot be plotted, 3 components are plotted to see if the clusters can be visualized. The resulting plot of the first 3 components is figure 11.

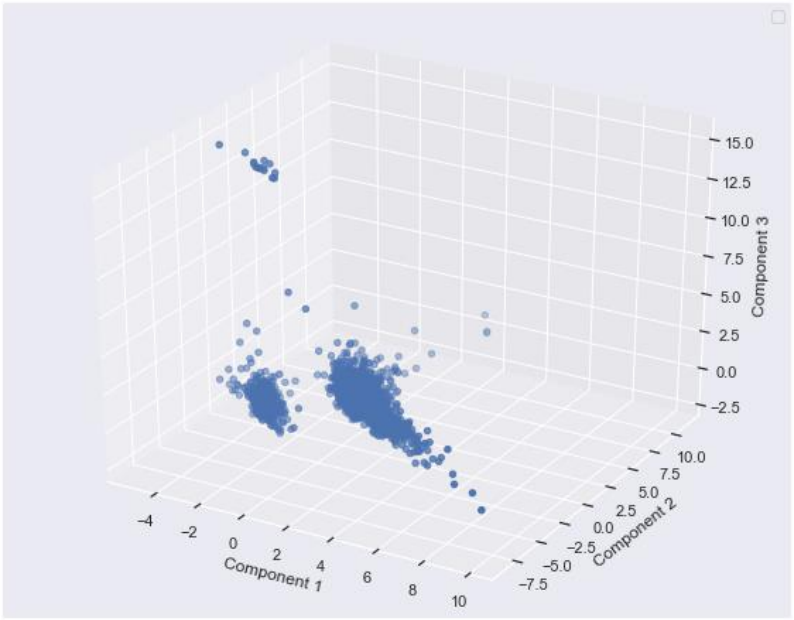

Figure 11: 3D-plot of the clusters.

Logistic Regression is used to build a model that can predict if a customer will bring revenue or not. To do this, the original data (including the 'revenue' variable is used).

Here, after encoding the categorical variable, the “Revenue” is made the target variable (y) while the other make up the independent variable (X).

Stratified shuffle split is then done to ensure equal representation of the target (y) variable is in the training and test sets of X and Y. The ratio of the stratified shuffle split is given in table 5.

Table 5: Number of rows in Training and Test Sets

Train		Test	
0	0.845209	0	0.845209
1	0.154791	1	0.154791

The training and test sets of X is scaled. Then the PCA is scaled using the X_train. This time around, k-means is not done. The number of components for the PCA model is 5 as seen in figure 12 and 13.

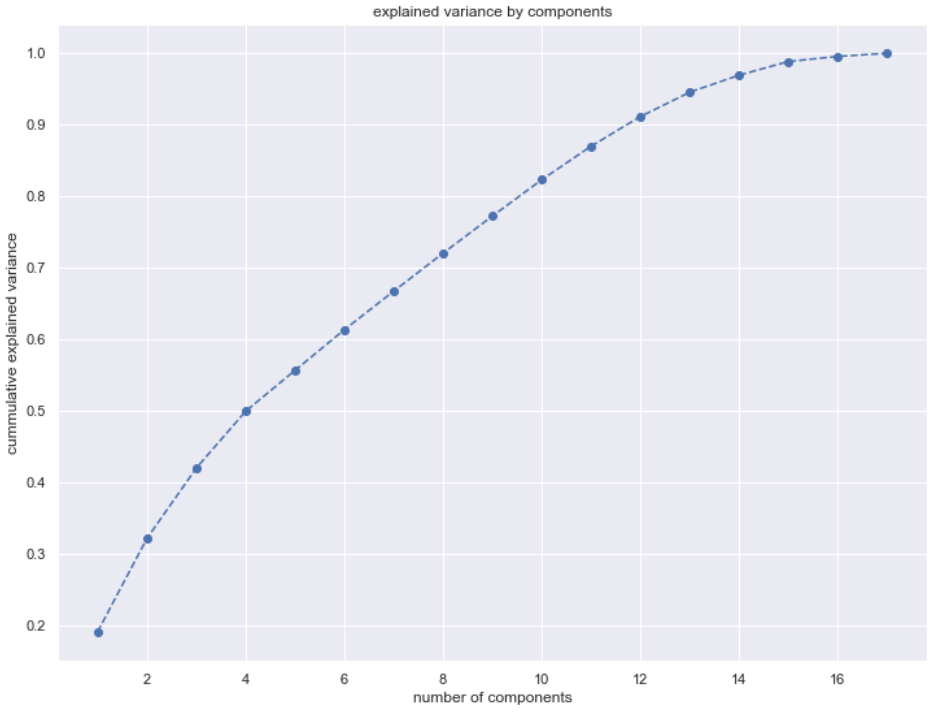

Figure 12: cumulative explained variance against number of components using the total (X_train) set

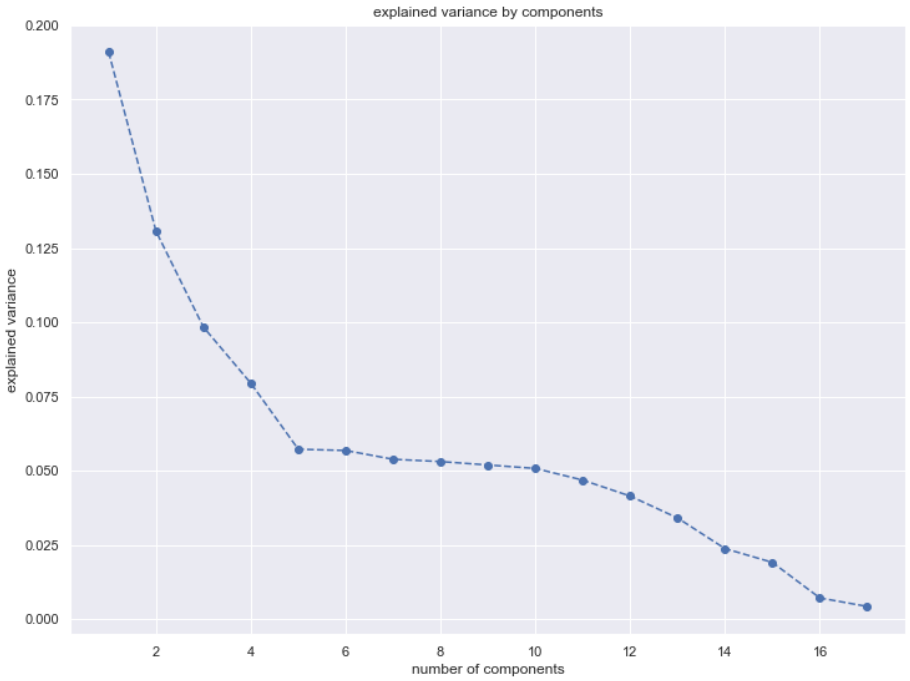

Figure 13: Explained variance of each component using the total (X_train) set

Since the second objective is to predict a categorical variable, the logistic regression, logistic regression with PCA and logistic regression with Kernel PCA are used. The accuracy, precision, recall, F1 and AUC scores are used for performance measurement of the models and results are depicted in Table 6.

Table 6: Performance Metrics of Models Used for Prediction

Performance Metrics	Scores		
	Logistic Regression	Logistic Regression with PCA	Logistic Regression with KernelPCA
Accuracy	0.878886	0.848263	0.837452
Precision	0.721429	0.712121	0.0
Recall	0.353147	0.111639	0.0
F1	0.474178	0.193018	0.0
AUC	0.865755	0.796181	0.317548

The Logistic Regression has the highest values across all accuracy metrics. However, comparing the logistic regression with PCA with Logistic Regression with KernelPCA, the Logistic Regression with PCA has a higher accuracy and performs better.

Observations/Suggestions

The PCA with K-means provided 4 clear segments among the customers who made purchase on the website.

The logistic model with PCA had an accuracy of 83.74% in predicting if a customer will make a purchase

To see the importance of each column in making prediction, ensemble methods such as random forest can be used.