# User interest information extraction using PCA and Kernel PCA

**Submitted to: Professor Dr. A Ben Hamza**

**By: Natasha Basutkar (Student Id: 40081017)**

***Abstract* – Humans are social beings; we love to socialise with our family and friends in our leisure time. However, in recent times the current generation are inclined towards technology and prefer staying indoors rather than doing outdoor activities. Taking this situation into consideration, there is a need to develop and improve the existing condition such that people start taking more interest in these outdoor activities. This problem can be recognised as a Classification task. The user interest information is captured to analyse the preferences of the public. Principal Component Analysis is applied to understand the critical features. This information can be used to improve the quality of public properties such as Places of Historic Importance, Educational places for children including Museums, Zoological parks, Sports complex, etc.**

***Index terms*— User Information reviews, holiday list dataset, principle component analysis, kernel PCA, Confusion matrix.**

## I. INTRODUCTION

User reviews given by 250 people during October 2014 is taken into consideration to understand the places of interest they would like to visit or spend their holidays[1]. Most of the users preferred visiting Beaches, Lakes, Rivers, etc. The second preference was given for entertainment such as Theatres, Exhibitions followed by Parks, Picnic spots, etc. However, only 2% people have shown interest in Sports which is fairly contrasting to the other activities that are close to 20%. This information can be effectively demonstrated using the Pie chart in Fig 1.

Considering the number of features in the dataset, Principal Component Analysis can be applied to get a better perspective of data by linearly reducing the dimensions of data. This allows us to focus on the significant features of the dataset.

Following PCA, Logistic regression technique is applied to define a good model that can make trust worthy predictions.

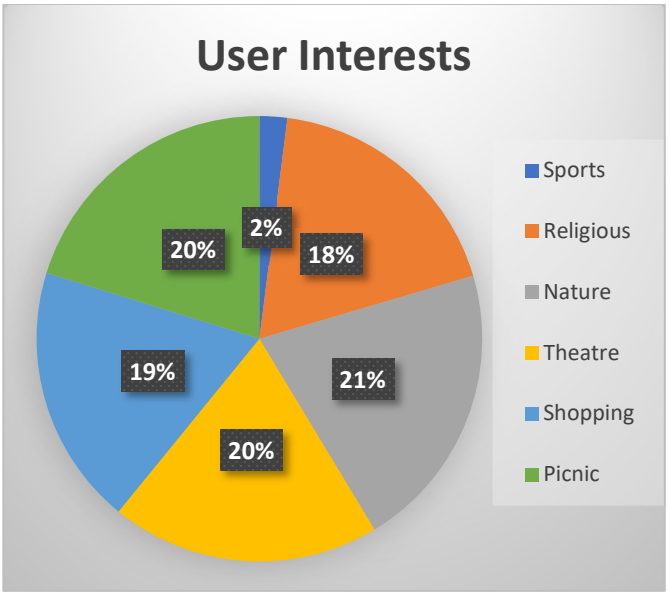


Fig 1: Pictorial representation of Holiday list dataset

## II. USER INTEREST INFORMATION DATA

The dataset considered for this project is taken from UCI Machine Learning Repository. The dataset consists of 249 reviewers of holidayiq.com. The Dataset contains 6 Attributes and a feature to list the User ID for each review. A sample of the first 20 rows of the dataset is posted below. Considering all the observations of the dataset, the following observations have been made.

- Mean of Nature attribute is greatest when compared to other attributes.
- The Sports attributes has the lowest deviation value in comparison to the other features.
- There is a huge contrast for the Minimum values of the listed rows, the lowest being 2 for Sports and the highest value is 61 for Picnic activity.
- Mean of Nature attribute is greatest when compared to other attributes.
- The Sports attributes has the lowest deviation value in comparison to the other features.

| Feature | Sports | Religious | Nature | Theatre | Shopping | Picnic |
|---|---|---|---|---|---|---|
| 1 | 2 | 77 | 7 | 69 | 68 | 95 |
| 2 | 2 | 62 | 7 | 76 | 69 | 68 |
| 3 | 2 | 50 | 9 | 87 | 50 | 75 |
| 4 | 2 | 68 | 7 | 95 | 76 | 61 |
| 5 | 2 | 98 | 5 | 59 | 95 | 86 |
| 6 | 3 | 52 | 10 | 93 | 52 | 76 |
| 7 | 3 | 64 | 8 | 82 | 73 | 69 |
| 8 | 3 | 54 | 10 | 92 | 54 | 76 |
| 9 | 3 | 64 | 10 | 64 | 54 | 93 |
| 10 | 3 | 86 | 7 | 74 | 74 | 103 |
| 11 | 3 | 107 | 5 | 64 | 103 | 94 |
| 12 | 3 | 103 | 6 | 63 | 102 | 93 |
| 13 | 3 | 64 | 8 | 82 | 75 | 69 |
| 14 | 3 | 93 | 5 | 74 | 103 | 69 |
| 15 | 3 | 63 | 8 | 81 | 78 | 69 |
| 16 | 3 | 82 | 7 | 75 | 75 | 82 |
| 17 | 5 | 59 | 13 | 103 | 54 | 86 |
| 18 | 5 | 56 | 12 | 108 | 56 | 85 |
| 19 | 4 | 85 | 6 | 111 | 65 | 72 |
| 20 | 5 | 114 | 8 | 65 | 114 | 102 |

Fig 2: Snapshot of dataset representing the features.

There is a huge contrast for the Minimum values of the listed rows, the lowest being 2 for Sports and the highest value is 61 for Picnic activity.

Similarly, the Maximum values are observed for Nature attribute resulting to 318 reviews.

| Feature | Sports | Religious | Nature | Theatre | Shopping | Picnic |
|---|---|---|---|---|---|---|
| Mean | 11.98795 | 109.77912 | 124.5181 | 116.3775 | 112.638554 | 120.4016 |
| Standard deviation | 6.616501 | 32.454115 | 45.63937 | 32.1327 | 41.562888 | 32.63334 |
| Minimum | 2 | 50 | 52 | 59 | 50 | 61 |
| Maximum | 25 | 203 | 318 | 213 | 233 | 218 |

Fig 3: Summary of holiday list data

The important features of the dataset can be effectively represented using the 3-D visualisation by plotting Nature, Theatre and Shopping on X, Y and Z-axis respectively. This graph tells us that the data is inconsistent and scattered across the dimensions. Hence, this data needs to be standardized before applying Principal Component Analysis.

The box plot for 6 features is plotted and shows that only Nature, Theatre and Shopping have outliers while rest of the datapoints are within the control limits. Another significant observation from the box plot is that Sports and Nature features are following close to normal distribution while Picnic attribute is negatively skewed and Religious attribute is positively skewed.

A bi-variate scatter plot is plotted to analyse the best set of features to explain relationships between two variables. We can note that Shopping vs Religious variables are heavily positively correlated. On the other hand, there is no correlation between Sports and all other attributes.

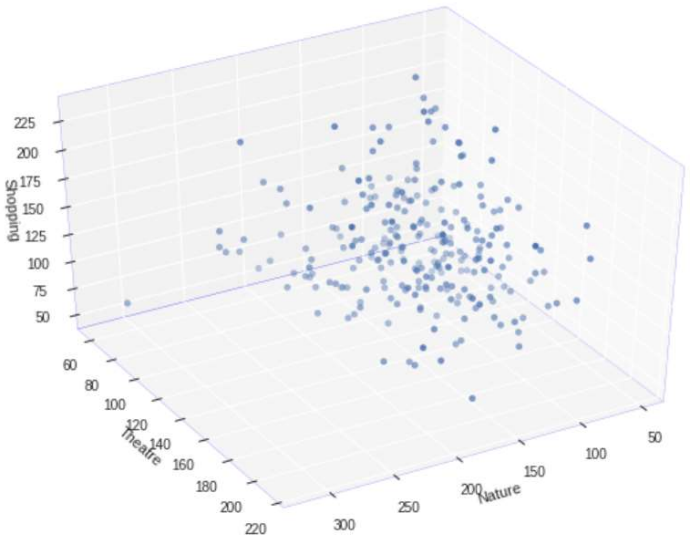


Fig4: 3D Visualisation of dataset

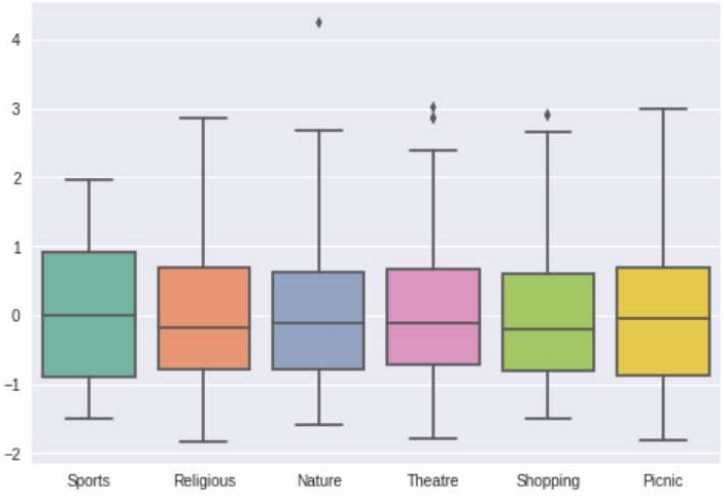


Fig 4: Boxplot for all attributes of the dataset

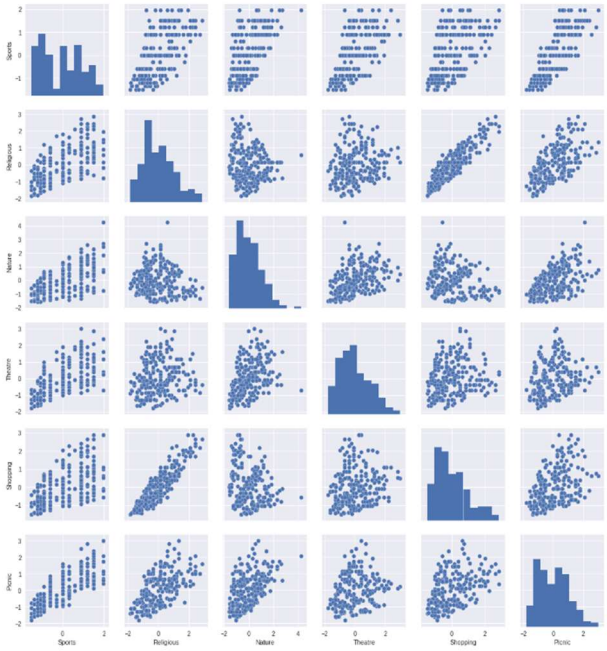


Fig 5: Bi-variate scatter plot for dataset

## III. PRINCIPAL COMPONENT ANALYSIS

PCA can be explained as Unsupervised linear dimensionality reduction technique to understand and reduce the number of dimensions in large complex datasets. Before applying PCA, the data is standardized, and the resulting data set has zero mean. This is then followed by plotting a Covariance matrix for the zero centered data[3].

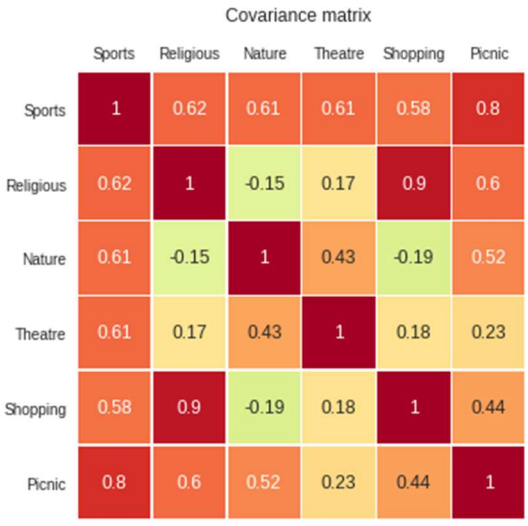


Fig 6: Covariance matrix of the dataset

The covariance matrix for our dataset shows how the attributes are associated with each other. We can see that Shopping and Religious values are almost to 1, meaning they are strongly correlated. This is followed by the correlation between Sports and Picnic features. Also we can see that Shopping and Nature are negatively correlated along with Religious attribute and Nature attributes.

**PCA Algorithm**

The following step is to apply the decomposition of the eigen values, as follows[2]:

S = A^AT;

Where S is the Covariance matrix, A is a p * p eigen vectors matrix and V is the diagonal matrix of eigen values. The eigen values and eigen vectors of the covariance matrix for our User Interests data is as follows:

```
A = [[ 0.53183266 -0.1630455  -0.024317    0.22006261  0.15779076  0.78527282]
     [ 0.43430306  0.45129164  0.01260212 -0.20374435  0.69709802 -0.2830203 ]
     [ 0.24426526 -0.64791327  0.25595772  0.44843906  0.20721922 -0.4593377 ]
     [ 0.29996214 -0.33808378 -0.80116173 -0.32486431 -0.12256709 -0.18248957]
     [ 0.4023048   0.47717358 -0.14286914  0.54265252 -0.49215375 -0.23099258]
     [ 0.46587866 -0.08924771  0.52102189 -0.55583677 -0.43471672 -0.07479956]]

Lambda [3.27409743 1.68817624 0.76514569 0.18444294 0.05860269 0.029535  ]
```

Fig 7: Eigen values matrix and Eigen vector matrix

The principal components are then calculated using the below formula,

Z=XA

Where X is the matrix with zero mean and the columns of the Z matrix are the actual orthogonal principal components. Hence the PC1 & PC2 components can be defined as follows.

**Z1 = 0.53183266Sports + 0.43430306Religious + 0.24426526Nature + 0.29996214Theatre + 0.4023048Shopping + 0.46587866Picnic**

**Z2 = -0.1630455Sports + 0.45129164Religious - 0.64791327Nature - 0.33808378Theatre + 0. 47717358Shopping - 0.08924771Picnic**

The principal components are plotted using a scatter plot for all the features of the dataset. We can see that few observations such as 236, 244 at the bottom of the chart and observations including 218,228,237 at the top right corner of the plot appear to be outliers.

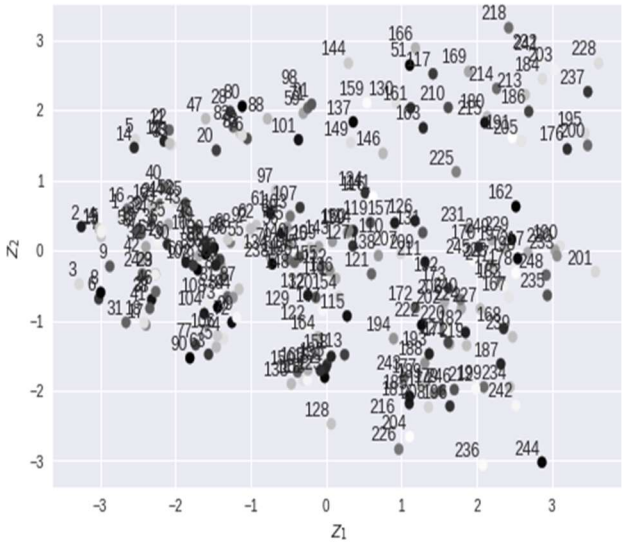


Fig 8: Scatter plot of PC1 & PC2 Score

The below plots describe the correlation between the PC1 & PC2 Co-efficient. From the below, we can say that Shopping and Religious attributes are closely related. While Picnic, Sports are located on the right while Nature attribute is located to the extreme left of the plot.

We can also infer that combinations such as Shopping - Religious, Picnic – Sports & Nature-Theatre are somewhat closely related. This is also evidently shown in the Z1, Z2 values. A striking feature is the difference between Religious and Nature attributes as the former has a positive co-efficient while the latter one has a negative co-efficient.


Fig 9: Scatter plots of PC1 vs PC2 Co-efficient

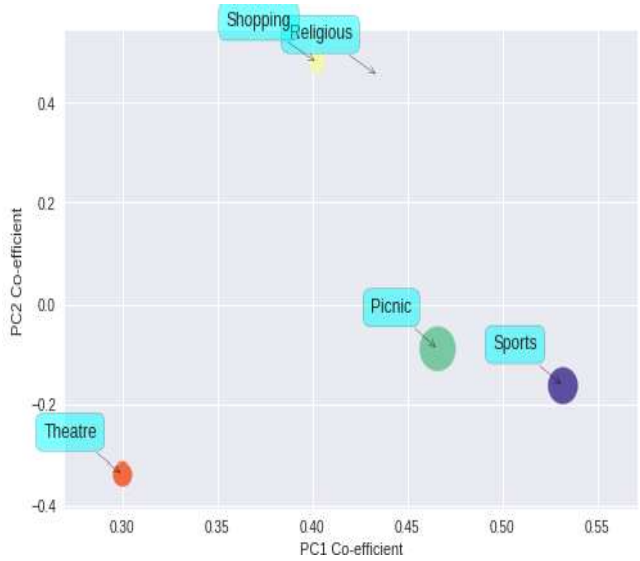


Fig 10: Scatter plot of PC1 vs PC2 Co-efficient

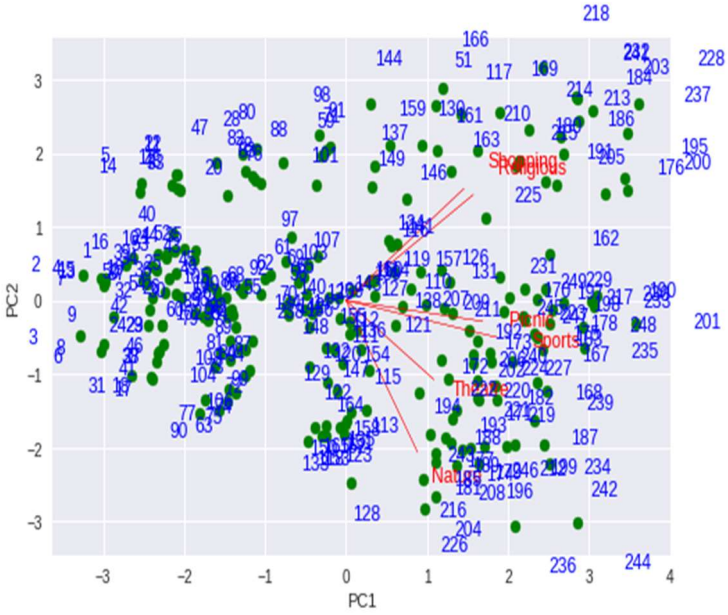


Fig 13: 2D Biplot

To understand the most important components from our dataset, we have used the explained variance and Pareto charts to differentiate the 'Vital most' from the 'Trivial many'.

A scree plot has been plotted to see the variance is distributed across the components of the data set. We can see that the first four components have a major impact followed by the rest of them. Numerically, the explained variance for the three PCs is given as below:

**ℓ1 = 54.5683%, ℓ2 = 28.1362%, ℓ3=12.7525%, ℓ4=3.074%**

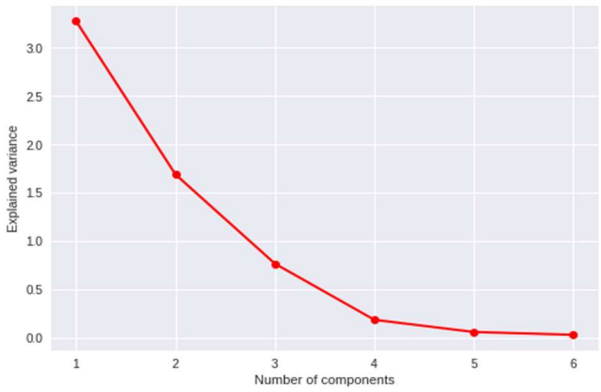


Fig 11: Explained variance plot

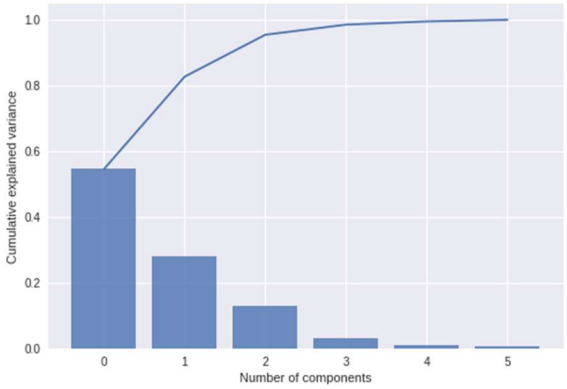


Fig 12: Pareto plot

From the above, we can say that the first two components equal to 82.7% of the explained variance in our holiday list dataset. While all the four components equal to 98.531% of the variance as per the scree plot.

The pareto plot ranks the cumulative variance against the number of components, starting from the highest to the lowest. As per the above chart, we can see that since the first 2 dimensions equal to more than 82.7%, so the minimum dimension required to explain our data is d=2 and the remaining trivial components can be eliminated.

The Bi Plot [Fig13] is used to represent the observations and variables of our dataset. We can see that the observations are plotted as points in Green color while the variables are plotted as vectors. Thus, biplots play a very important role in Principal component analysis. The following observations can be made from our Biplot for the Holiday list data:

- The picnic variable is closely aligned to the X-axis indicating the importance of contribution this variable is making to the axis dimension.
- Shopping and Religious attributes are closely correlated to each other since the cosine of the angle seems to be very less.
- Similarly, Picnic and Sports vectors are also strongly correlated with each other.
- We can see that observations from 1 to 90 present in the second and third quadrants are very close to each other and hence represent the observations which have similar values.

The below heatmap is plotted to understand the correlation between eigen value after performing PCA analysis. We can see that Sports and Picnic are very close to each other and hence are colored Red.

Similarly, the Nature and Theatre values are weakly correlated, this is evident from the correlation value being negative and hence is colored Green. We can see that attributes such as Shopping and Theatre are normally correlated without exhibiting strong relationships among them. This is also evident from the Biplot above.

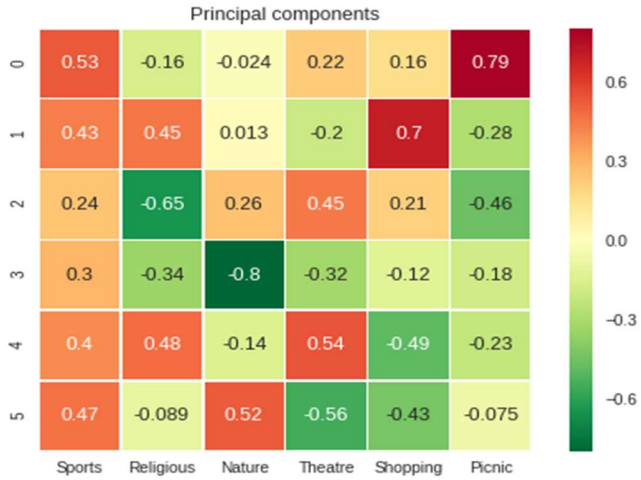


Fig 14: Heatmap for Principal components

## IV. KERNEL PCA

Kernel PCA is a non liner dimensionality reduction technique used for data that is not linearly separable. Hence, to solve linearly inseparable data, we use the non linear technique to reduce the dimensionality of the datasets and understand the output model of the most significant features.

**Why use Kernel PCA**

As described in the below figure, the data in the left graph the datapoints are not aligned in a straight line. In this case, PCA cannot comprehend and reduce the dimensionality of this data from 2-D to 1-D[4].

However, Kernel PCA is capable of understanding this non linear curves and represents that data in nearly one-dimensional space. The Kernel PCA has also proved to be efficient for Polynomial kernels too. It can be applied for various datasets from Classification, Regression or Clustering algorithms to develop models based on the training sets[6].

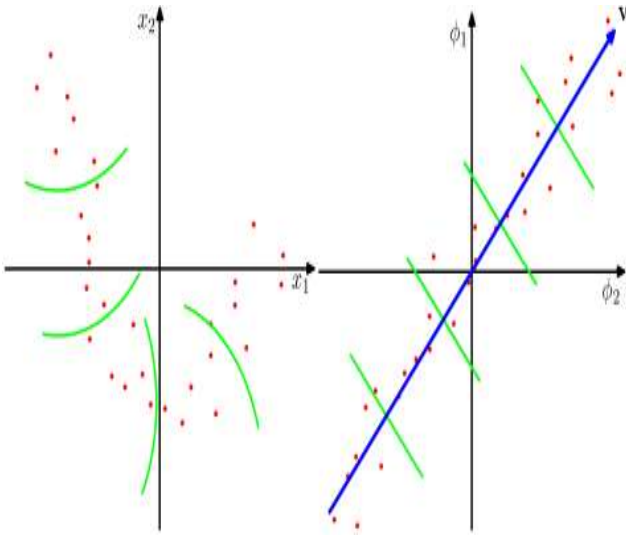


Fig 15: Kernel PCA Algorithm[9]

To apply the Kernel PCA to our User Interests dataset, we start by identifying the significant features of our dataset. We have realised that Shopping and Picnic are good candidates on which the Kernel PCA can be applied.

**Step 1:** Identify the Kernels. We have identified Shopping and Picnic as appropriate Kernels for our dataset.

**Step 2:** We then construct the normalized Kernel matrix K, of dimension m x m for our data.

**Step 3:** The next step is to find the Eigen values and Eigenvectors for the Kernel matrix. They can be represented as $\lambda_j$ , $a_j$.

**Step 4:** The following set of features are represented for any datapoints of the dataset.

$$y_j = \sum_{i=1}^{m} a_{ji} K(\mathbf{x}, \mathbf{x}_i), j = 1, \ldots m$$

**Step 5:** Each feature, $y_j$ is a coordinate of $\varphi(x)$ and lies along the feature space axes vectors, represented as $v_j$.

**Step 6:** Since vj are orthogonal vectors, the projection of $\varphi(x)$ onto the higher dimensional space is represented as $\Pi\varphi(x)$.

**Step 7:** With this data, we plot the dot-products on the training and test datasets. However, the errors in the models is represented as below:

$$\|\phi(\mathbf{x}) - \Pi\phi(\mathbf{x})\|^2$$

The User Interests Information is divided into Training and Test datasets. Of the 249 rows, 186 observations accounting to 75% of the data is allotted to training datasets while the remaining 25% of the data is allotted to test datasets.

The Kernel PCA and Logistic Regression functions are applied to the datasets and the below graphs are plotted.

**Points to note:**

The training set model for Shopping vs. Picnic has more errors when compared to the test set model. This infers that the predictions made for the test set are strong and have resulted in less number of errors.

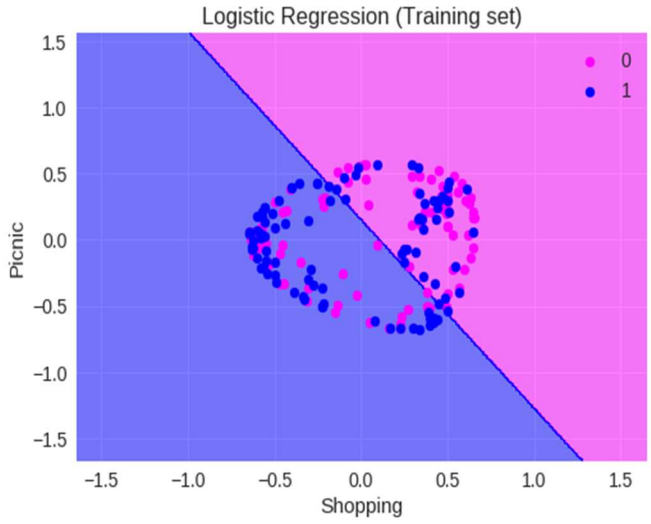


Fig 16: Scatter plot for Training dataset

There are approximately 30 error spots in the Shopping training set model which have reduced to 11 errors in the test set model. Hence there is a reduction of 63.34% of errors in the test dataset.

This gives us sufficient evidence to confirm that our output model is a good quality classification model.

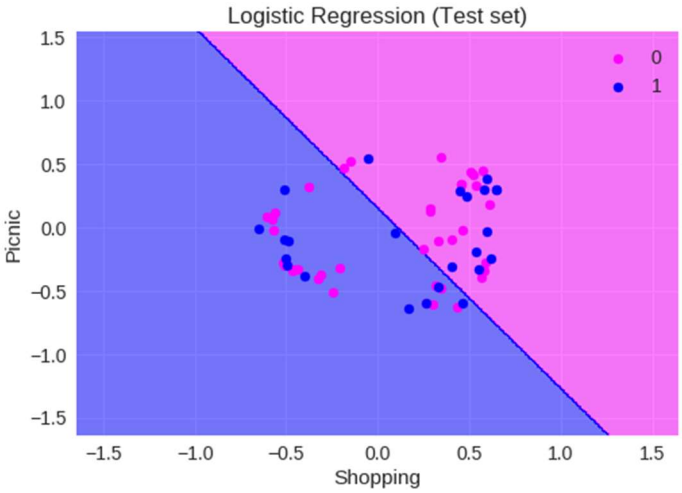


Fig 17: Scatter plot for Test dataset

**Confusion Matrix to verify Performance of our model:**

An error matrix, commonly known as Confusion matrix is a table used to explain the performance of the classification models. It gives us a quick and accurate review of the performance of our User Interests testing data, where the actual values are unknown. The confusion matrix is divided into 4 quadrants briefly explained below[8]:

**True Positives (TP):** It occupies the First quadrant in the matrix and takes value 1(True) when the actual value of the data point is True and the predicted class is also True.

**True Negatives (TN):** It lies in the 2nd quadrant and takes value when the actual is False and the predicted is also False.

**False Positives (FP):** It lies in the 3rd quadrant and takes value when the actual value is False and the predicted is True.

**False Negatives (FN):** It occupies the 4th quadrant and takes value when the actual value is True and the predicted is False.

There are a number of performance metrics that can be derived from the confusion matrix as described as follows[7].


Fig 18: Confusion matrix

**Accuracy:** The accuracy of the predicted model is given by the formula and the accuracy for our model is resulting to 49.3%. So we can infer that the model is in a mediocre state and needs further improvement to improve the accuracy.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

**Recall:** Recall can be defined as the number of items that have been correctly identified as positive among all true positive values.

Recall value for our data model is calculated as 48.7%, since the recall value is in the mid-range, the actual and predicted classes need to be more properly recognised.

$$\textit{Recall} = \frac{TP}{TP + FN}$$

**Precision:** Precision can be calculated by dividing the number of positively classified items by total number of predicted positives. The precision value for our model is accounting to 61.3%.

This shows that quite many positive samples have been missed but the ones predicted positive are certainly positive.

$$Precision = \frac{TP}{TP + FP}$$

**F – Measure:** Considering the Precision and Recall values to predict the accuracy, another simple and effective parameter that represents both of them is the F-Measure. The F-Measure works on the principle of Harmonic mean as it is considered to give more consistent results when compared to Arithmetic mean.

Consequently, if either of the Recall or Precision values is small, the Harmonic mean is a number closer to the smaller figure. Thus, this helps the model in giving an appropriate score. It can be defined as below:

$$F\text{ - measure} = \frac{2*Recall*Precision}{Recall + Precision}$$

The F-Score for our User Interests data model is close to 54.27% while the Arithmetic mean for our model would have been 55.0%. In our case, we don't see a big contrast in the values since there is not much difference between Recall and Precision values.

## V. CONCLUSION

In this project, we have considered the User reviews and tried to analyse the holiday preferences. To do this, we have performed linear regression using PCA in the first part and in the latter phase, we have used Kernel PCA to identify the most significant features of the dataset. From the test dataset, we can say that since there are few errors in the plot, the prediction model is considerably a good one. However, from the F-measure we can conclude that the model is considerably in the mid-range and requires improvement to increase the performance and quality to a higher level.

References:
[1] http://archive.ics.uci.edu/ml/datasets/BuddyMove+Data+Set
[2] https://www.khanacademy.org/math/linear-algebra/alternate-bases/eigen-everything/v/linear-algebra-finding-eigenvectors-and-eigenspaces-example
[3]http://www.stephacking.com/principal-component-analysis-pca-python-import-libraries-import-data/
[4] https://github.com/jasminelatham/Dimensionality-Reduction/blob/master/jbk_pca.ipynb
[5] https://seaborn.pydata.org/generated/seaborn.pairplot.html
[6] https://stats.stackexchange.com/questions/94463/what-are-the-advantages-of-kernel-pca-over-standard-pca
[7] https://towardsdatascience.com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-3-classification-3eac420ec991
[8] https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
[9]Kernel PCA Diagram Figure, taken from Bishops Pattern recognition and Machine Learning textbook