95-851: Data Science for Product Managers
Section A3

Final Project Report: Deloitte Data Analysis

Tabassum Kazi
Kaiqi Bian
Steffi Nazareth
Prithvi Poddar

Introduction

In 2012, Steier&Corp had an advertising giant, Inmobi, as their client. Inmobi was looking to improve upon using different types of platforms and applications as their advertisement delivery mechanisms. Utilizing Deloitte's 2011 digital demographic survey data, Steier&Corp conducted various levels of analysis and modelling in order to provide adequate recommendations.

The challenges we are targeting are:

- What platform/device should Inmobi utilize?
- What channel on said devices is the optimal channel?

Through our exploratory analysis, we gain insight on our demographic distribution, and clean our data in preparation for preprocessing and modelling. Consequently, we use K-modes clustering as our unsupervised learning method to identify the clusters in which various platforms lie. This also provides further demographic insights as well. Lastly, we gauge the optimal smartphone applications to utilize as an advertisement channel through predictive analysis conducted through Support Vector Classification. Through this supervised learning technique, we are able to predict, given an individual, the probability they will use a given type of application. Given that probability, we may discern if that is a viable target application. Moreso, the feature importance from our classification model underscores for each type of application the most important questions, and in turn, more demographic insight on application usage.

Inmobi, or advertising giants in general, may ultimately benefit from our 'problem-solution'. Customer and market segmentation allows for the effective allocation of marketing resources and employment of devices, tools, and strategies.

Scope of Our Data

The data we are working with is Deloitte's Digital Demographic Survey data from 2011. Through this survey, we gain insight into the technology consumption patterns across various media, devices, and demographics. The data contains 2131 records of survey results and 198 columns. The columns represent the answer choices to questions served in the questionnaire. Some key assumptions we made about this data are:

- At first glance, there are many missing values. However, many questions and multiple choice answers were dependent on prior question answers. Hence, not everyone was necessarily asked the same questions. This discrepancy has to appropriately be handled in order to not lose meaning.
- Research was done to identify the ideal columns (devices and applications) that provide the most relevant insight to our research goal

Preprocessing & Exploratory Data Analysis

Firstly, renaming columns to readable and accessible names was our priority.

Next, in order to take care of the abundance of NULL values, we made these key assumptions:

- If someone does not have any other children living in their home (QNEW1), the NULL values in the children's age groups (QNEW2) were replaced with 'No'. Because, if no children live in the household, there are no children in the separate age groups as well.
- If someone does not own a smartphone (Q8), the NULL values in applications used on smartphones (Q22) are replaced with 'No.' If one does not own a phone, they do not use these apps regardless.
- If a device/subscription/entertainment method was not ranked (Q11, Q36, Q37), NULLs were replaced with a -1.
- To account for the fact that the data is mixed categorical and numerical data, one-hot encoding of categorical variables was implemented.

Demographic distributions shed light on the unbalanced nature of the dataset. This can be referenced in our jupyter notebook file and will be further discussed in our recommendations.

How people rank devices:

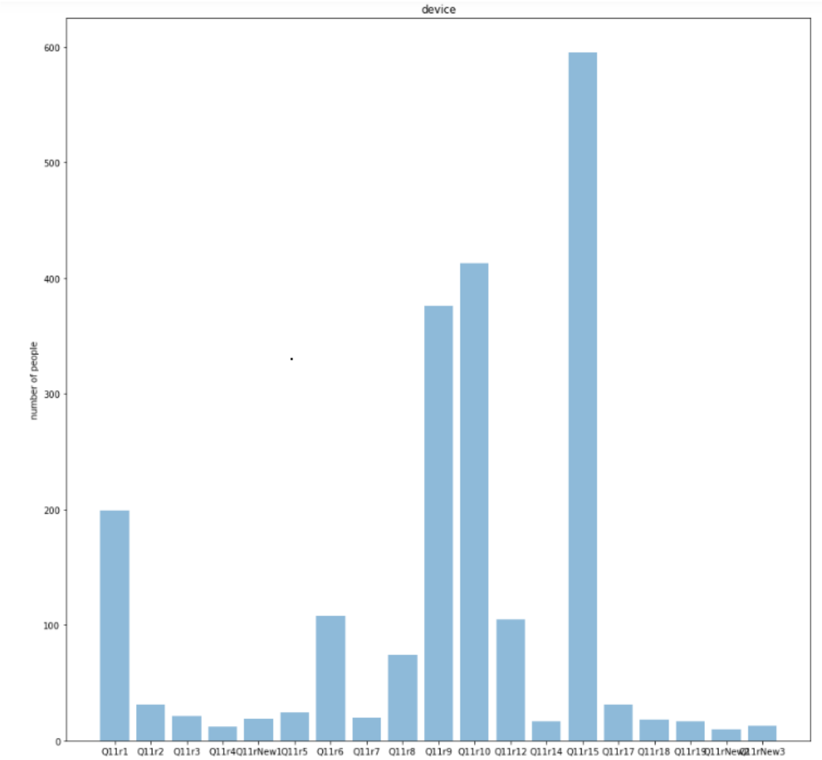

Figure 1: Most Valued Devices

The top four #1 ranked devices, in order, are: Smartphones, Desktop Computers, Laptop Computers, and Flat Panel Televisions.

Our target platform for potential advertisements overlap with these most-valued devices. We target smartphones, desktops, laptops, and tablets. While tablets were not ranked #1 most valued, they are of meaning due to their similar nature to smartphones in housing applications that can deliver advertisements to users.

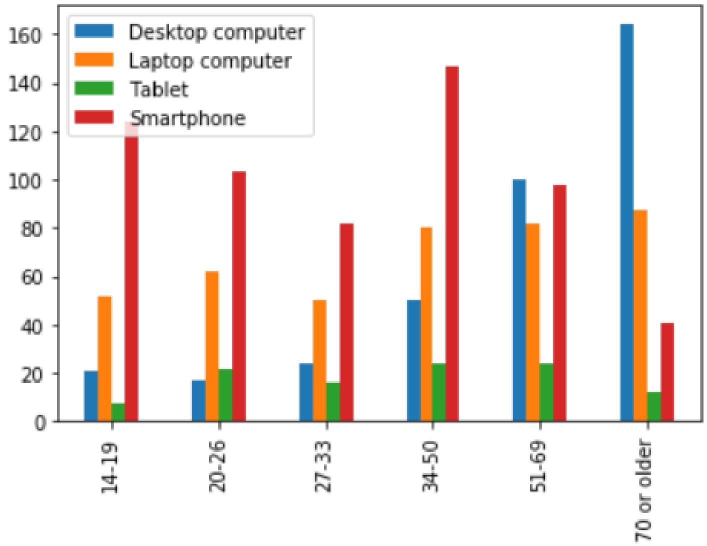

Figure 2: Device Ranking Grouped By Age

Assessing the demographics that own these target devices further, we notice that while people ages 14-50 primarily rank smartphones as their #1 device owned, people over 70 years old primarily prefer desktops.

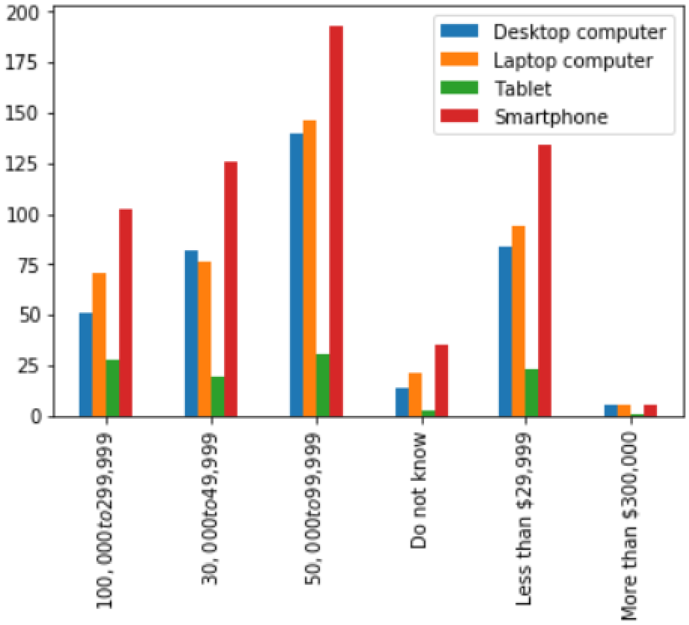

Figure 3: Device Ranking Grouped By Income

Amongst all income brackets, smartphones seem to be ranked as the most valued device across the board. This is interesting, as smartphones were just making their place in the market during this time.

More demographic breakdowns by device ownership and ranking can be viewed in our jupyter notebook file.

We also analyzed the summary statistics of the primary device used to pursue certain forms of entertainment activity. People reported the percentage of time they use a smartphone, television, tablet, or desktop/laptop computer to view TV shows, movies, or sports. For example, people spend, on average, watch TV shows on their smartphone 10.4% of the time. Whereas, they watch TV shows on an actual television 44.6% of the time.

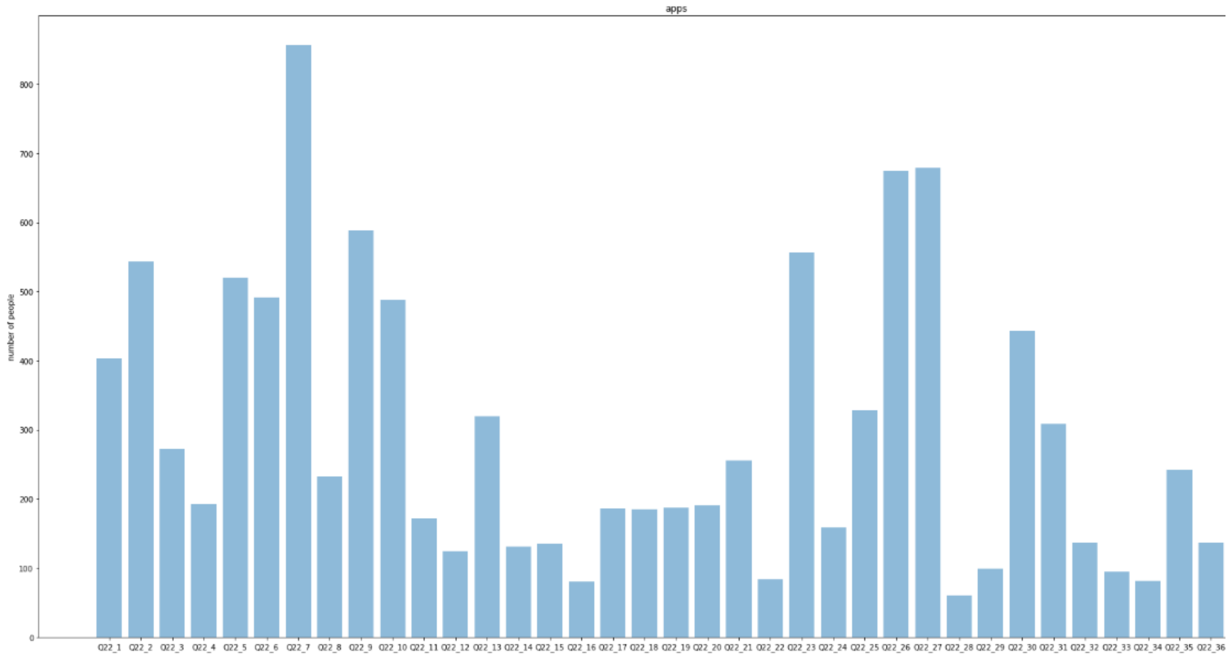

Figure 4: Most Frequently Used Apps

Similar to device rankings, we also viewed the usage frequency of smartphone applications. The top 3 most used applications on smartphones, in order, are: social networks, weather, and internet browsers.

Our target applications do not represent the top 3 applications used on smartphones, they do represent the top 3 applications used that are also primary advertisement channels. These are comprised of: music streaming, gaming, and video streaming applications.

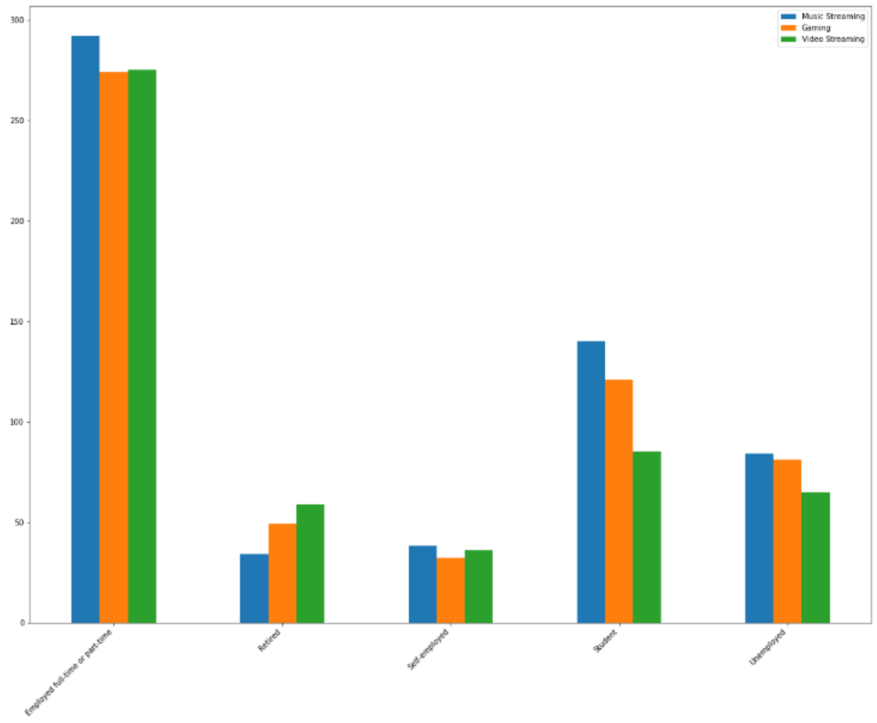

Figure 5: Frequently Used Applications Grouped By Employment

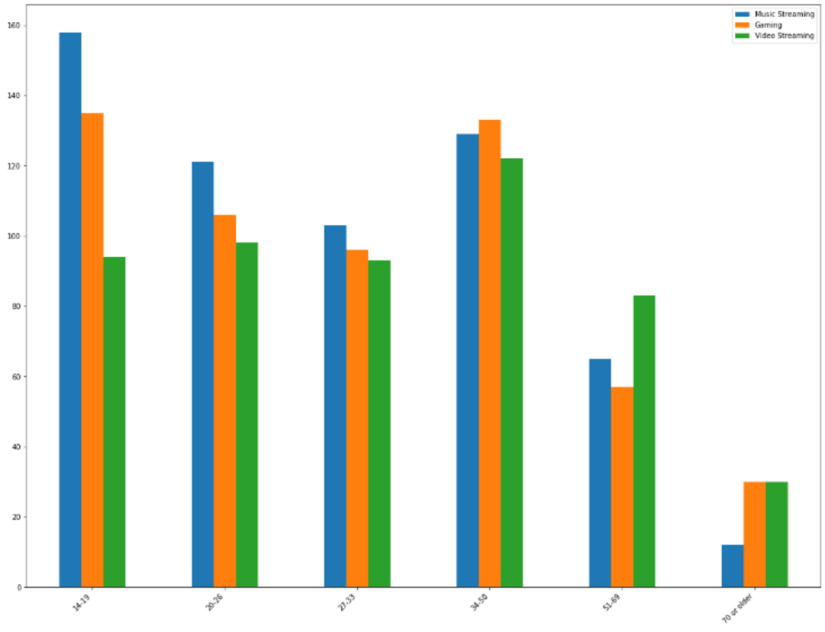

Figure 6: Frequently Used Applications Grouped By Age

Application usage exhibited less variation and insight than did the device ranking distributions. Nevertheless, this level of exploration of our dataset was essential to understanding it thoroughly. More demographic data regarding application usage can be found at our jupyter notebook file.

Unsupervised Learning: K-Modes Clustering

K-modes clustering was used to classify characteristics of questionnaire respondents in accordance with our target devices.
In order to find the ideal number of clusters, we used the elbow method.

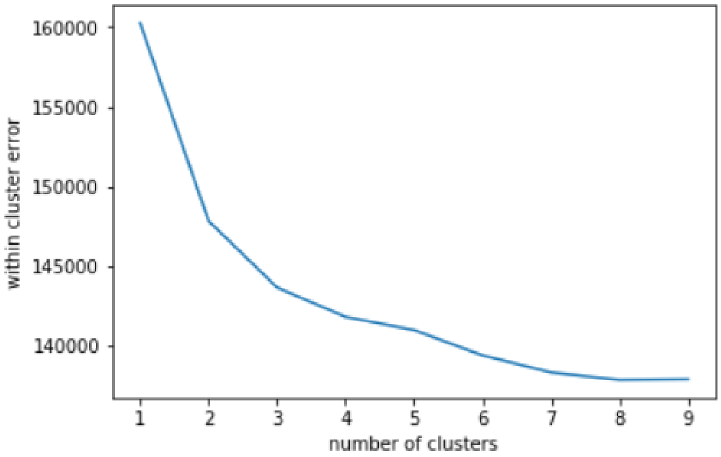

Figure 7: Elbow Method

From this, we decided that the optimal number of clusters to use was 4 clusters.
Below exhibits the distribution of records in each cluster.

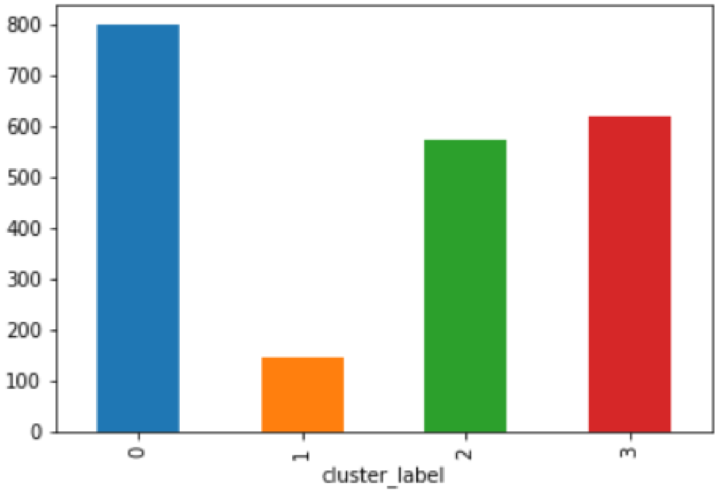

Figure 6: Cluster Distribution

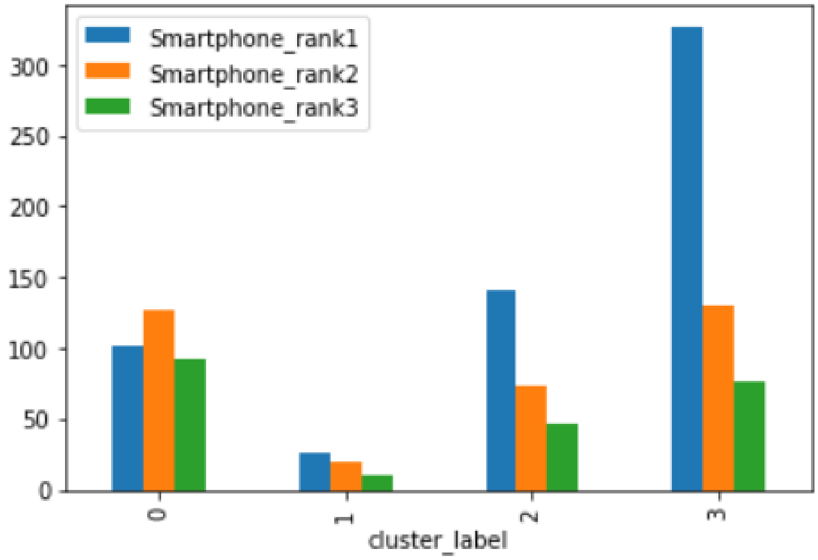

Figure 9: Clustered Smartphone Rankings

From this, we gather that the people who value smartphones as their number one device primarily lie in cluster 3. Further analysis of the demographic of this cluster may provide information. Distribution of our other target devices can be referenced in our jupyter notebook file. Our findings regarding the other devices align with our results from our EDA.

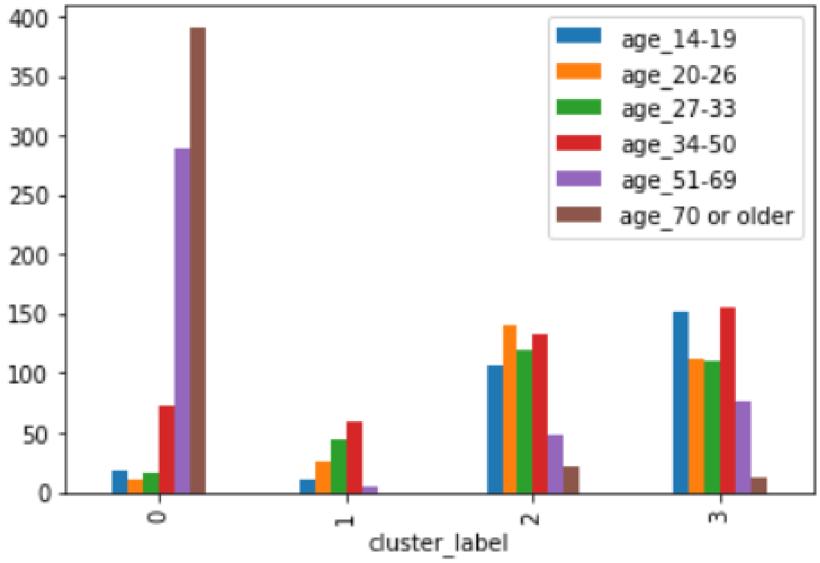

Figure 10: Age Across Clusters

The age distribution in cluster 3 encompasses quite a wide range. Almost each age bracket is represented except the 70 or older bracket, which lies primarily in cluster 1. This solidifies our findings during our EDA, in which almost every age group except the elderly valued smartphones the most.

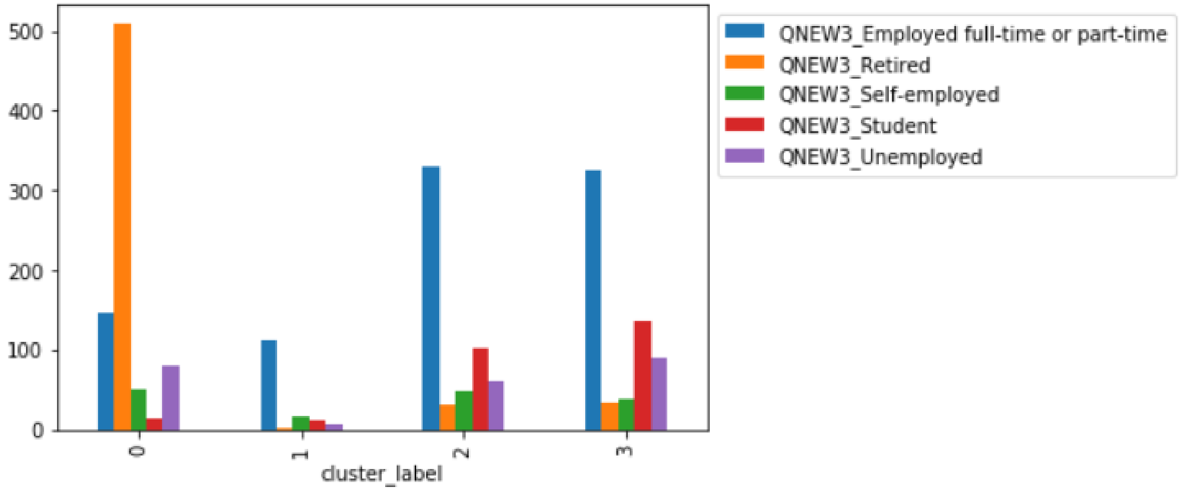

Figure 11: Employment Across Clusters

Similarly, we viewed employment distribution across our clusters. The majority of cluster 3 appear to be either part- or full-time employees. This can be explained by our wide range of representative ages as well. Further cluster features are analyzed in our jupyter notebook file.

We also chose to observe the application usage distribution across the four clusters. This proved less insightful, as cluster 3 already proved to value smartphones the most, so they would use applications on their smartphones the most as well. We can note that music streaming applications are slightly more frequently used than the other two types of applications.

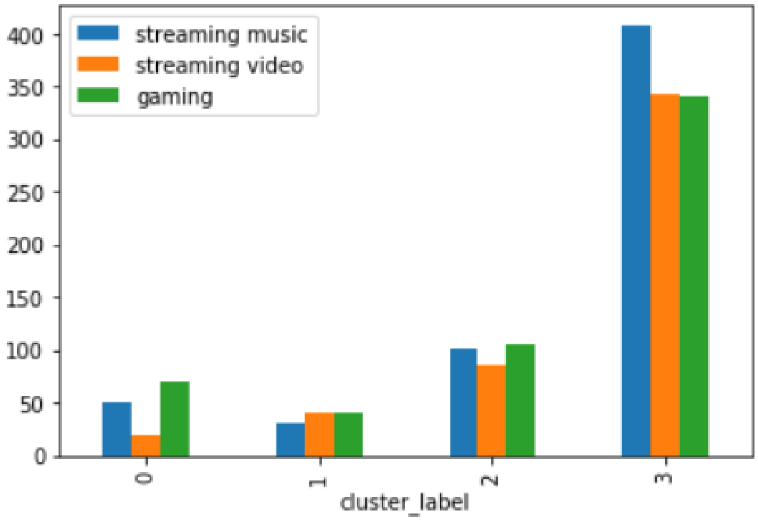

Figure 12. Target Applications Across Clusters

Supervised Learning: Support Vector Classification

We used Support Vector Classification in order to predict the type of user who prefers which type of our three target applications. We performed three different prediction models, each differing in the target variable. The three target variables are our three target applications: music streaming, gaming, and video streaming. The models were trained on 75% of our full dataset, and tested on 25%.

Music Streaming Application Model: This model has an accuracy score of 76.9%.

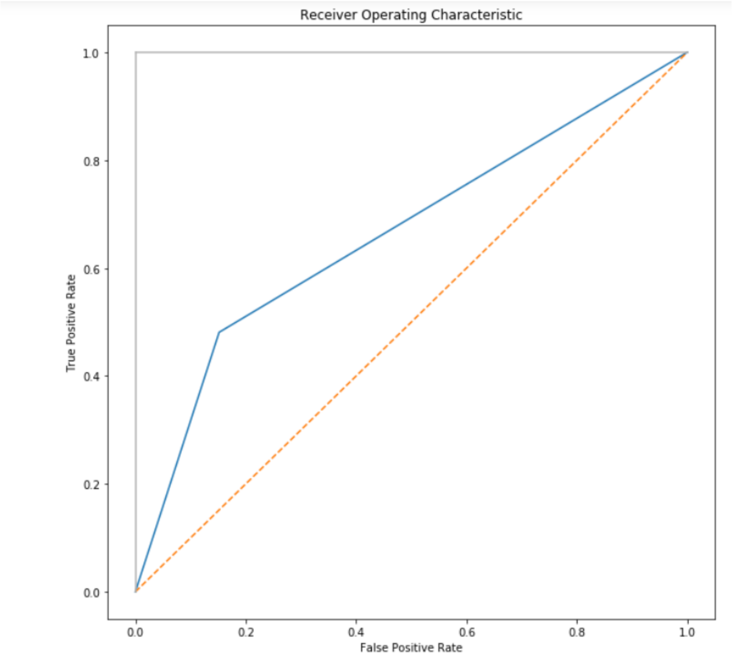

Figure 13: ROC Curve, Music Streaming Applications

Its AUC score was 71.1%.

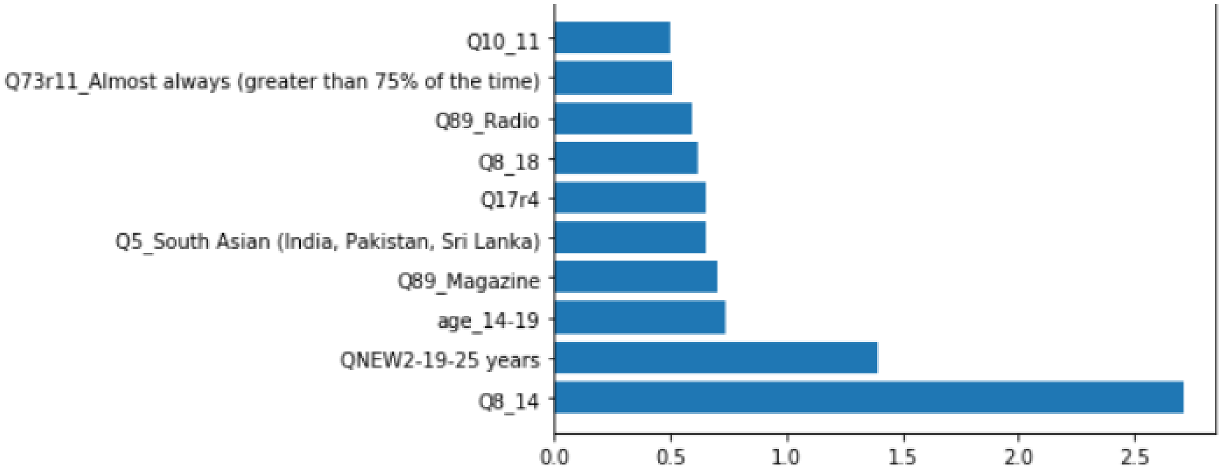

Figure 14: Feature Importance, Music Streaming Applications

As expected, the most important feature is the ownership of the smartphone, as this is a model based on smartphone applications. It is interesting to note that the next most important feature is the age group of 19-25 year olds.

Gaming Application Model: This model has an accuracy score of 77.3%.

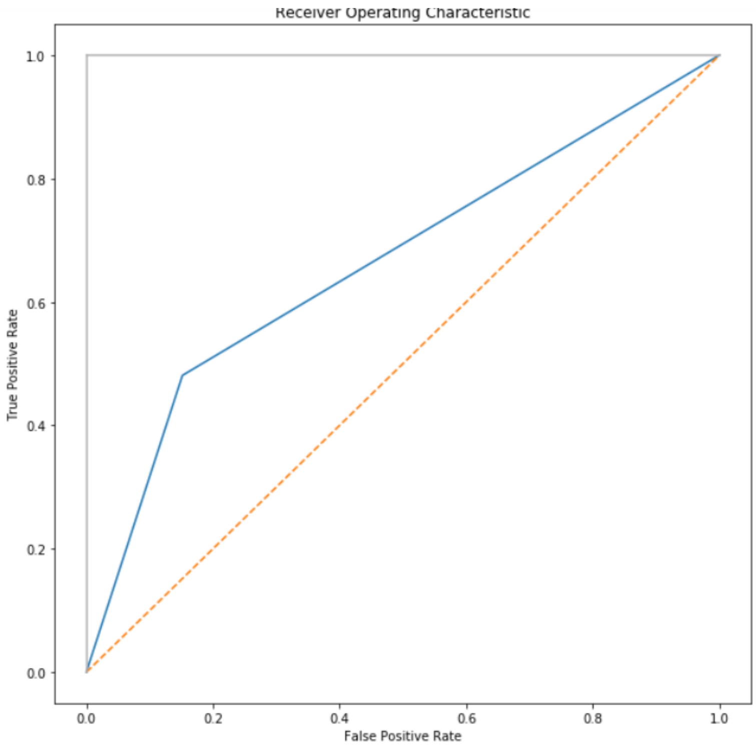

Figure 15: ROC Curve, Gaming Applications

Its AUC score was 69.5%.

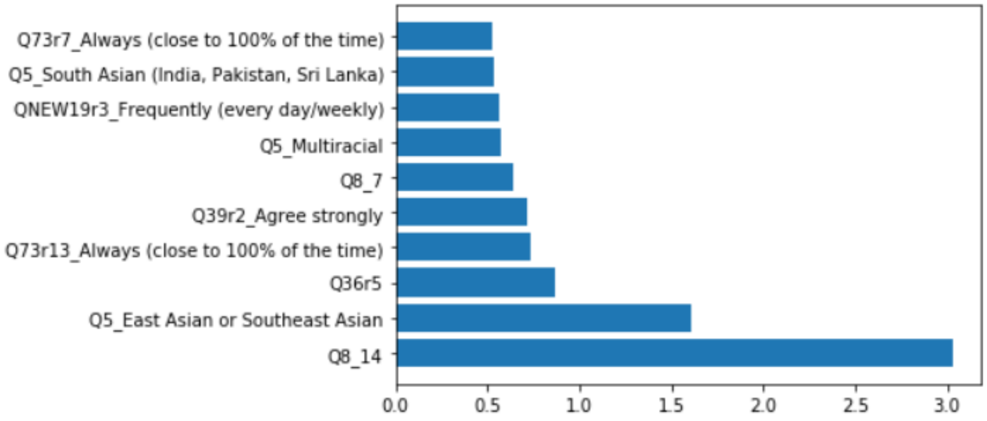

Figure 16: Feature Importance, Gaming Applications

Especially of interest, the next important feature is Q39r2, for which this type of person agrees strongly. This question asks whether this person is willing to provide personal information in exchange for targeted and personalized advertising. Interestingly, Q73r13 asks what users do even while watching television, and these users answered that they play games close 100% of the time.

<u>Video Streaming Application Model:</u> This model has an accuracy score of 76.2%.

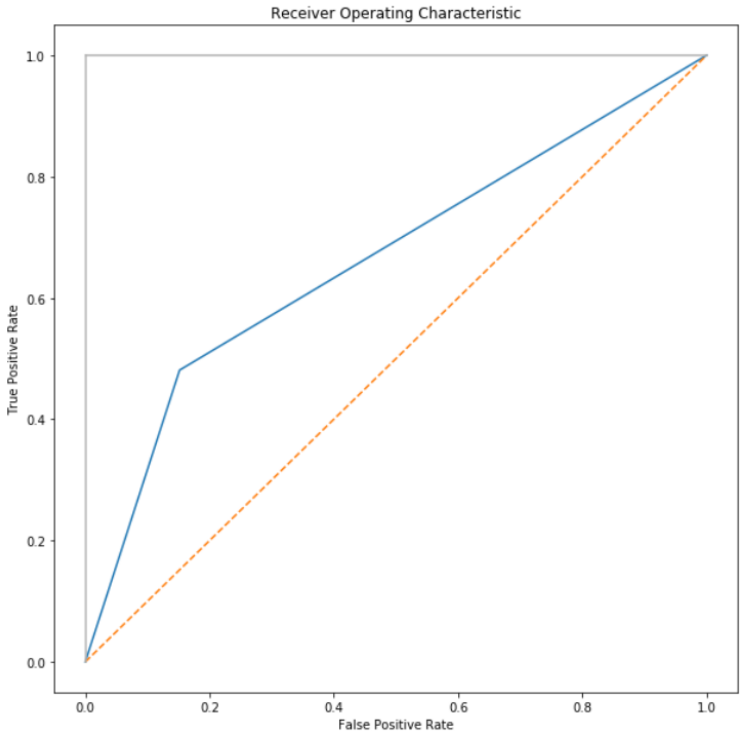

Figure 17: ROC Curve, Video Streaming Applications

Its AUC score was 66.5%.

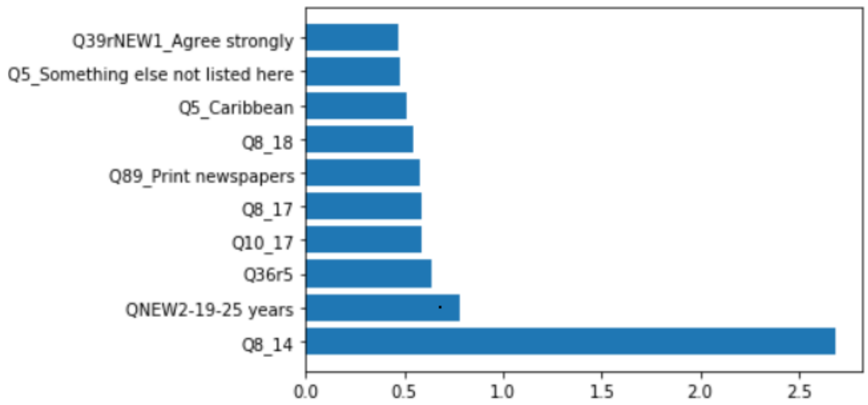

Figure 18: Feature Importance, Video Streaming Applications

More demographic inferences can also be made here as well.

Results & Recommendations

As per our analysis, smartphones seem to be the best platform in terms of valuation by consumers. Of applications on smartphones, gaming application users are likely to even provide additional information for further targeted advertisements. This segment is comprised of the Asian population represented by this survey. Ultimately, these models should help identify by predicting which type of the three advertising channels an individual suits the most.

Further recommendations include:

- Clustering did not adequately provide information on the other platforms we hoped to study. The next steps may consist of including questions regarding desktop/tablet application usage as well.
- The data is highly unbalanced, with a large representation of elderly and white records. This is very characteristic of surveys. However, it is not ideal for analysis purposes. Measures towards preventing bias revolving around surveys should be employed further.
- How to cope with those who have subscriptions to get rid of advertisements
- Further explore for different target variables
- Further explore across time trends