Observing Online Streaming User Value in iflix in Sri Lanka

[1st]M.K.G.Sandeepani
Department of Computer Science & Engineering
Faculty of Engineering
University of Moratuwa
sandeepanimkg.20@uom.lk

[2nd] D.M.R.M. Dissanayake
Department of Computer Science & Engineering
Faculty of Engineering
University of Moratuwa
dissanayakedmrm.20@uom.lk

***Abstract*—The purpose of this analysis is to observe online streaming Sri Lankan user value in iflix which is a video on demand service. This study can be used to identify the audience well. Understanding the audience plays a major role in content creating, decisions making within the company, packages offering etc. The study contains descriptive analysis, diagnostic analysis and predictive analysis for the dataset. The potential customer base for iflix service was identified after analyzing the dataset.**

Keywords—video on demand, iflix, data analytics, data science, user profile segmentation, regression analysis

I. INTRODUCTION

Nowadays, people are busy with their day to day work. So, they do not have time to go to a theatre to watch a film and get relaxed. People are willing to get done everything easily and at their fingertips in this modern world. They prefer to follow the same method when it comes to entertaining themselves. There, video on demand service comes into the world.

Video on demand is a technique which allows users to get video entertainment without a traditional device or a typical broadcasting schedule. Iflix is such a video on demand service which focuses on emerging markets. The content of iflix consists with a large selection of TV shows, movies, news, live sports etc. The iflix user can access to its content at anytime from anywhere over the internet using the technology called 'video streaming'. The user can access iflix when he is in online. If he wants to use the service offline, it is also possible by downloading the content. Moreover, iflix service is accessible by a smartphone, laptop, tablet or TV. People prefer iflix because of its attracting and enjoyable service. Therefore, it currently has more than one billion consumers in all around the world.

There are iflix users in Sri Lanka as well. The user profiles of iflix audience guide towards so many insights which help to understand the audience very well. It would be impossible to fine tune the strategy of iflix without understanding the audience. Therefore, we are investigating the iflix users in Sri Lanka and this report is based on the analysis of iflix user profiles in Sri Lanka. It may address the following research questions.

Research Question 1: How do the users access the iflix service?

Research Question 2: Users in which age group highly use the service?

Group No:01 Below 18 yrs

Group No:02 Between 18 yrs and 25 yrs

Group No:03 Between 25 yrs and 30 yrs

Group No:04 Between 30 yrs and 40 yrs

Group No 5: Above 40

Research Question 3: How the users have been distributed by geography?

Research Question 4: How the users have been distributed by gender?

Research Question 5: What kind of account type do the users highly use?

Research Question 6: What is the potential customer base?

II. METHODOLOGY

Dataset was about the user profiles in iflix in Sri lanka and it was found from an R&D company in Sri Lanka. It contained data for the attributes such as age, gender, data usage, access device etc. It had to get preprocessed before using it as there were some missing values in it. There, the mean value was used for continuous variables and the mode was used for categorical variables in the dataset.

Descriptive analysis, diagnostic analysis and predictive analysis were done on the dataset. It was helped to collect insights on data. After that, a model building process was done to propose a plan for a new customer by identifying the potential customer base from above analysis.

The study is based on basic data science concepts. Python libraries such as sklearn, matpotlib and keras were used in the study.

III. DESCRIPTIVE AND DIAGNOSTIC ANALYSIS

The analysis is based on several attributes such as age of the user, gender, average monthly bill amount, accessing device details, account type, data usage etc. The above research questions were addressed by using the attributes of the data set and finally some valuable insights were introduced.

First, we analyzed the data usage by district wise to get a general idea about the data usage of the customers in different districts. It is displayed below in Fig.1 and Fig.2.

The first research question is, 'How do the users access the iflix service?'.

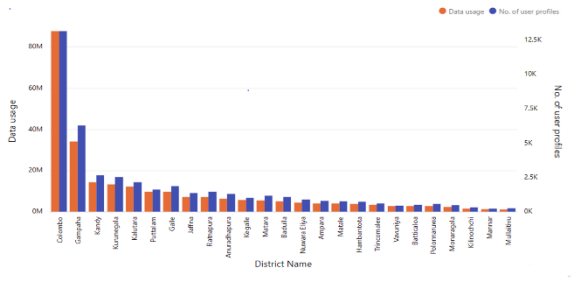

Fig.1. No. of user profiles and Data usage by District

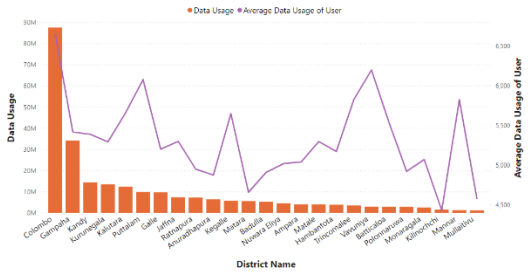

Fig.2. Comparison of Average Data Usage of Users by District

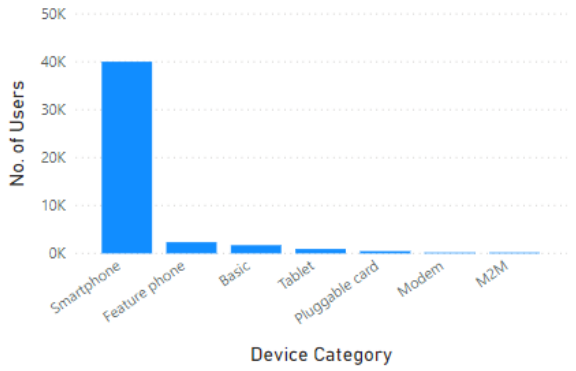

Fig.3. User Distribution by device category

According to the dataset, iflix users in Sri Lanka use various devices such as smart phones, tablets, feature phones etc. and they prefer various brands such as Samsung, huawei, apple, htc etc. to get the iflix service. Accessing device and the device brand gives a hint about the user's income. So, it can be used in content creating, offering packages in iflix. The result of the descriptive analysis on user distribution over the device and the device brand is displayed in Fig.3.

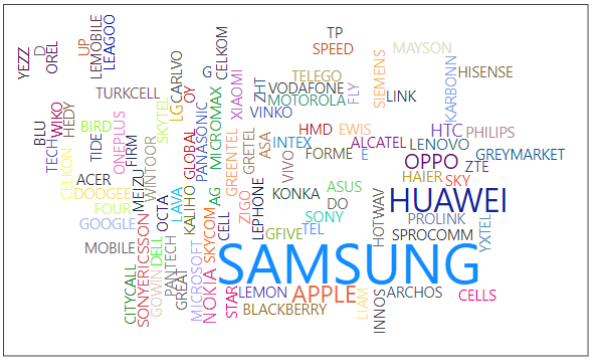

Fig.4. User distribution by device brand

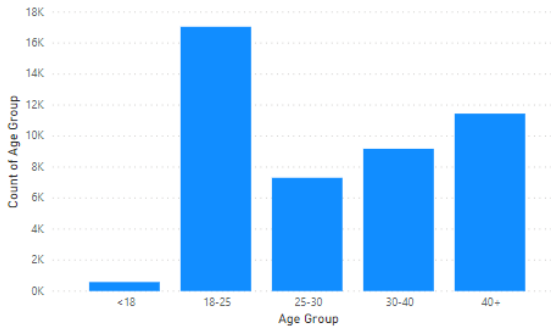

Fig.5. User distribution by age group

Output of descriptive analysis implies that, most of the users prefer smart phones to get mentioned service. Today, everyone tends to use a mobile phone and they prefer to buy a smart phone as it can be used with a lot of functionalities. Smart phone is affordable for many people and portable with compared to laptops and tablets. Therefore, the widely used accessing device for iflix may become the smart phone.

The analysis was done further to identify the device brand. The result is shown in Fig.4. Device brand indirectly connects with the user's income. According to the results it was identified 'samsung' as the most distributed brand among the users. After, doing a market research about several brands it was visible that 'samsung' is the widely used smart phone in Sri Lanka. People have the image as 'samsung' is user friendly, higher quality and price is affordable for most of the Sri Lankan smart phone users. It is one reason for having higher number of iflix users who are using 'samsung' brand. Brands such as htc, phillips, zte are used by a smaller number of people to access iflix with compared to Samsung.

The second research question is, Users in which age group highly use the service. According to the data set, the iflix users in Sri Lanka belong to different age groups. So, the ages of the users were categorized into five age groups in the research and analyzed the user distribution according to the age group. The Groups are mentioned in the introduction section. The output of the analysis is displayed in Fig.5.

According to the analysis, most of the users belong to 18 - 25 yrs age group. People who are in age 18 -25 yrs are considered as young adults in society. Normally, young adults are familiar with technology and they like to enjoy life as they take most of the things easy. Further, they generally do not have problems such as family problems. It may be the main reason for belonging most iflix user profiles into the age group of 18 - 25 yrs.

The number of user profiles are smaller in the first age group (below 18 yrs) with compared to other age groups. People who are in below 18 yrs are considered as children. They are less facilitated with smart phones, tablets, and money. It may be the reason for above result.

In order to come up with more insights, the analysis was further done by comparing age groups over to the data usage. The result is displayed in Fig. 6.

It is clearly seen in the graph that data usage is higher in second age group (between 18 to 25yrs)

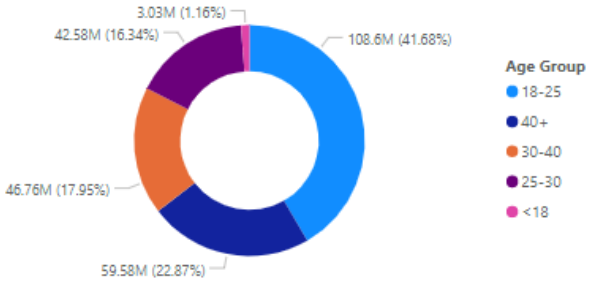

Fig.6. Data usage by age group

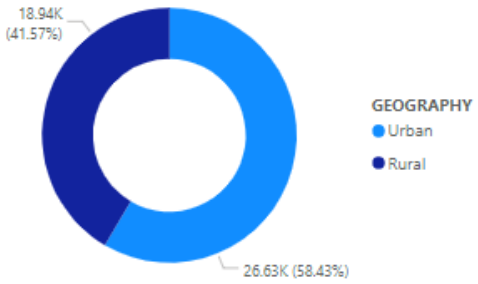

Fig.7. User distribution by geography

The third research question is, 'How the users have been distributed by geography'. According to the data set, users are in urban area or in rural area. The respected analysis is shown below in Fig. 7. It can be clearly seen in the analysis that most of the users are distributed in urban area.

Then the analysis was done for data usage by geography. The outcome is displayed in Fig. 8. There, it was clearly seen that the data usage is higher in urban area than in rural area.

One reason for not having more iflix user profiles in rural areas in Sri Lanka may because of weak signal strength of the service provider within the particular geographical area. So, people avoid using the services such as iflix when the connection gets disturbed because of the signal strength.

The fourth research question is, How the users have been distributed by gender. The relevant analysis is displayed in Fig.9. According to the diagram majority of the users are males. Female percentage who use iflix is very less. In order to observe more insights by doing the diagnostic analysis, we analyzed the gender distribution over to the district. It gave the output which is shown in Fig. 10. According to that most of the users are males who live in Colombo and Gampaha districts. Colombo is the capital of Sri Lanka and urbanization in Colombo and Gampaha districts is higher than in the other districts. Therefore, males in districts far from the capital generally spend their free time by farming, spending time with family etc. Because of that they may not prefer entertaining services such as iflix.

Further analysis was done to get an idea about user distribution with the preferred device brand. The output is shown in Fig. 11. According to the analysis it was shown that high cost brands such as 'Apple' are less used by people in districts far from the capital.

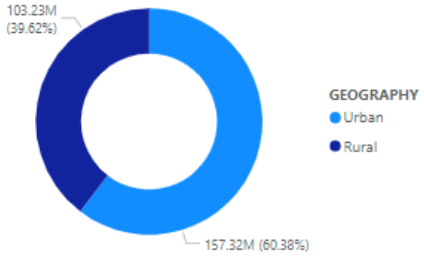

Fig.8. Data usage by Geography

Next research question is, 'What kind of account type do the users highly use? '. Analysis output shows that most of the users tend to use 'prepaid' account type. The relevant graph is shown in the Fig.12. When comparing prepaid account features with the postpaid account features it is significant that, it is able to pay only for the usage in prepaid account. Therefore, people prefer prepaid connections.

Further, postpaid account type indicates the stability of the customer. According to the analysis done for the previous research questions it was identified that the most iflix users in Sri Lanka are males in the age group 18 to 25 yrs. Generally, that category is not stable in life. That may be a reason for getting the answer as 'prepaid account type' for the above research question.

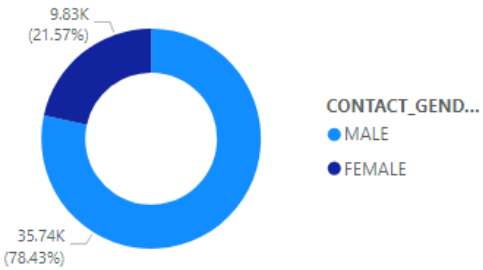

Fig.9. User distribution by gender

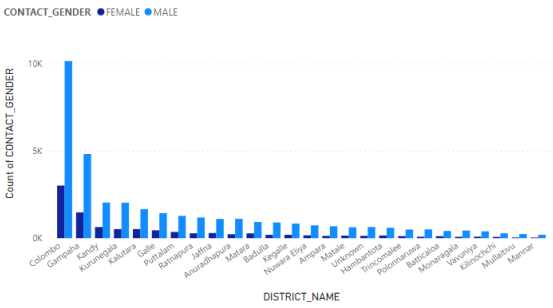

Fig.10. User distribution by gender and district

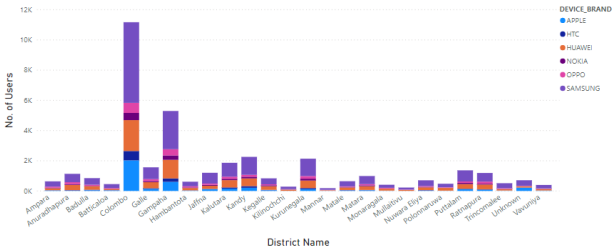

Fig.11. User distribution of brand preference in the districts in Sri Lanka

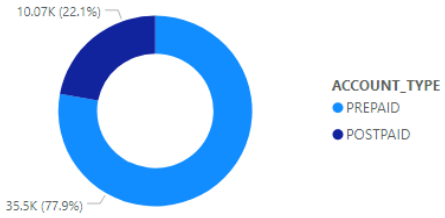

Fig.12. User distribution by account type

IV. PREDICTIVE ANALYSIS

Data clustering is the task of dividing a dataset into subset of similar items. Applying user profile clustering to iflix dataset generates group of similar user profiles. These groups are called user profile segments. Similar users have strong, measurable relationship among them. The strength of a relationship between two or more user profiles can be quantified as similarity measure; A mathematical function computes the correlation between two user profiles. The result of that computation, called similarity values, essentially compare a particular user profile to all other profiles in the dataset. Those other users will be either more similar or less similar in comparison to that specific user.

In order to that, User profile segmentation has been done to cluster profiles using Machine learning approach. It is shown in Fig.13.

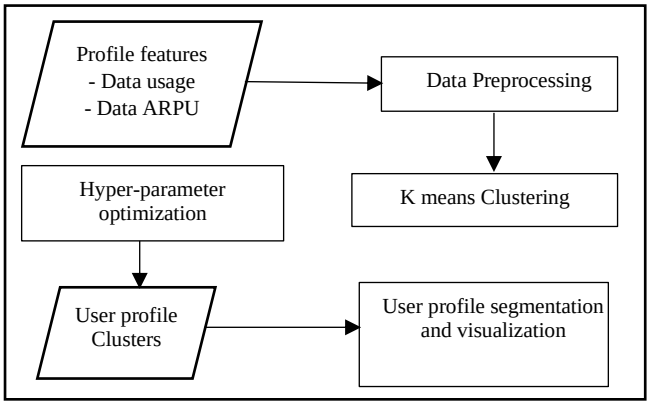

Fig.13. User Segmentation

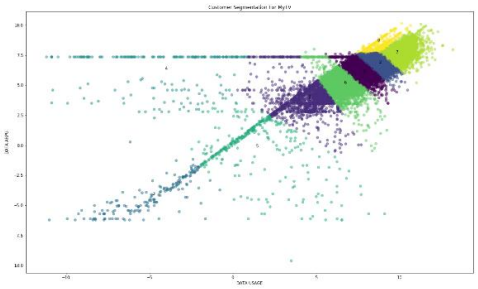

Fig.14. User profile Segments

In this module we segment user profiles based on data usage and data ARPU. Mean of each attribute has been used to replace empty values and equal or below to the zero values of each attribute as preprocessing stage. We used k-means clustering for our task. We used randomized optimization to identify optimum properties for k-means algorithm. The optimization parameters are below:

- "init": ['k-means++', 'random']
- "n_init": sp_randint(1, 20)
- "max_iter": sp_randint(300, 600)
- "n_clusters":sp_randint(1, 10)

And optimized model for clustering is: 'init': 'random', 'max_iter': 417, 'n_clusters': 9, 'n_init': 16 in K-means clustering model.

Fig.14 shows user profile segmentation. Each point represents the unique user profile by the axis of Data usage and the data ARPU. the color of each data points shows which user profile segment is belongs the unique user profile. Based on the user profile segmentation Cluster 2 and 7 and 8 can be **most potential clusters** (It is straightforward that Clusters 2,7 and 8 have highest Data usage compared to other clusters, also these two clusters are comparatively better than other clusters for Data ARPU)

We can use this analysis for the market strategy of iflix. For example, to identify the potential customer base. When it comes to iflix, the most important pieces of information is to know about potential customers For example to identify who they are, age of them, gender, geography, device category etc. It is also good to know about their data usage, how they make revenue for the company etc. So, we can do the predictive analysis of their expectations of business, and what they think of the business. Knowing the customers in depth is a key, in order to be a highly successful business. Moreover, identification of the customers who do not generate more revenue for the company is also important aspect. From that the marketing team can analyze the attribute of that segment and do the marketing strategy to make them as potential customers in future.

Furthermore, user profile clustering can also help the business to identify, learn, or predict the nature of new users Especially, how new user can be linked with making predictions. For example, in pattern recognition, analyzing patterns in the data (such as Data usage patterns in particular Geography or age groups) can help business to develop predictive analytics. In this case, predicting the nature of future data items that can fit well with established patterns.

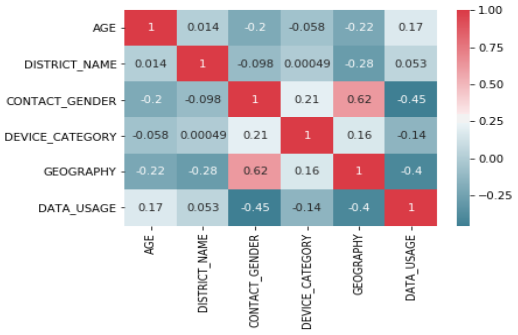

Fig.15. Correlation Matrix

Algorithm	R2 value
Linear Regression	0.3921
Decision Tree Regression	0.1454
SVM	0.2014
Neural Network	0.1121

Table 1. Accuracy of the models

V. PREDICTION OF DATA USAGE FOR A NEW USER

This experiment has done for predict the data usage of new user when he/she joins with iflix. This regression prediction has been done by using user attributes such as age, district name, contact gender, device category and geography. We can gather those type of data by the new user when he/she registers to the iflix app. The target variable is data usage. Fig.15 shows the correlation between features and the target variables.

Feature variables are both nominal and ratio. In here the reason for using regression model is to build for the prediction of data usage for a new user since target variable is continuous. We have done this experiment for multiple regression algorithm and neural network algorithm to optimize the model accuracy. 70% of the dataset has been used as training the model and 30% has been used as test data. Before train the model, categorical features have been converted to the numerical values by using one-hot encoding method. R2 value has been calculated to evaluate the accuracy of each algorithm tested. Table 1 shows the results;

According to the above result, the Linear Regression model is most suitable for this scenario among tested algorithms. But the performance of the model is not that much better for the prediction of real-world scenario. The reason for that is lack of features which has enough correlation with the target variable. But as the result accuracies can be increased by applying available strategies such as feature selection and hyperparameter optimization etc.

The purpose of this experiment is to suggest for iflix, this approach is very practical and relevant to current business competition in attracting new customers to their product. This is very useful prediction for both customer and company side as well. From that they can come up with the most suitable data plan for the new customers when doing the registration for the iflix app. It can be applied as marketing strategy of the company itself.

VI. CONCLUSION

The analysis was done to observe online streaming user value in iflix user profiles in Sri Lanka. Main purpose of the study was to identify the potential Sri Lankan customer base for the company. In order to do that descriptive analysis, diagnostic analysis and predictive analysis was done for the dataset.

ACKNOWLEDGMENT

Acknowledgement goes to the staff Dr. RTUthya and Dr. Shehan Perera in department of Computer Science & Engineering, Faculty of Engineering of University of Moratuwa for guiding this study.

REFERENCES

[1] Herwina Rosnan, Nuraisyah Chua Abdullah, “The Internationalization of Online Video Content Industry” 25th Silver International Conference. 14-15 July 2018.

[2] Shreya Tripathi, Aditya Bhardwaj, Poovammal E, Approaches to Clustering in Customer Segmentation, International Journal of Engineering &Technology, 7 (3.12) (2018) 802 -807.

[3] Chinedu Pascal, Constance Kalu, Simeon Ozuomba“Application of K-Means algorithm for Efficient Customer Segmentation: A Strategy for Target Customer Service” International Journal of Advanced Research in Artificial Intelligence · October 2015.

[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049417/