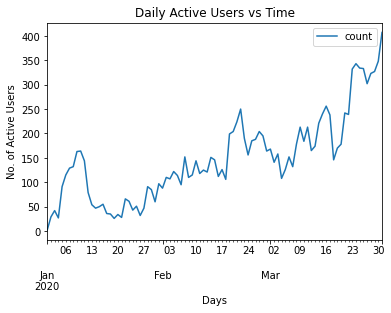
***Data Analytics***

Q0. Data

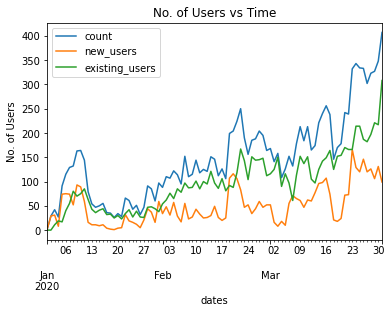
Raw data consists of dates and User Ids logged in to an entertainment platform on the date. There are 91 days from January 1st 2020 to March 31st 2020, February consists of 29 days; There are 13527 total logins to the platform and 4594 unique users logging in over the 3 months. There are 2596 One time Users. January 1st 2020 has the least number of logins with 2 while March 31, 2020 has the most number of logins at 407.

Q1. Daily Active Users (DAU) varied across time



Daily Active Users is the number of users per day. From the above graph you can see an initial increase in the first half of January after launch of the platform then a decrease in the second half of January. There is an overall increase in the number of active users at the end of March with few sudden spike increases in the users. As you can see lowest number of users is on January 1st while highest is on 31st March with value of 407

Q2. Proportion of DAU contributed by new users vs old users



Here in the graph, the Blue line indicates the total number of Users per day while the Green line is for existing users and Orange is the new users.

From the graph you can see that the number of new users is almost always less than the number of existing users except in the first half of January where the number of new and existing users are similar.

Q3. Retention Metrics

Daily Retention is the calculated by the number of users on day x since the user’s start day.

Example : Say User 1 start date is jan 3rd and for User 2 Start date is Jan 5, User 3 is Jan 6, User 4 is Jan 3rd. For January 3rd - 7th,

| User Id, Date | 1 | 2 | 3 | 4 |
| --- | --- | --- | --- | --- |
| Jan 3 | 0 | - | - | 0 |
| Jan 4 | 1 | - | - | 1 |
| Jan 5 | 2 | 0 | - | 2 |
| Jan 6 | 3 | 1 | 0 | 3 |
| Jan 7 | 4 | 2 | 1 | 4 |

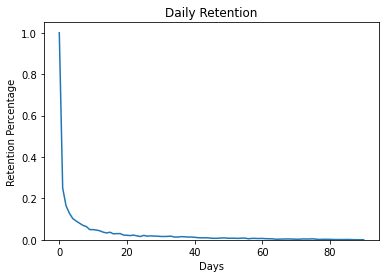
So for days since start, we have

| Days since start | User Ids | Count | Daily retention |
| --- | --- | --- | --- |
| 0 | 1,2,3,4 | 4 | 1 |
| 1 | 1,2,3,4 | 4 | 1 |
| 2 | 1,2,4 | 3 | 0.75 |
| 3 | 1,4 | 2 | 0.5 |
| 4 | 1,4 | 2 | 0.5 |

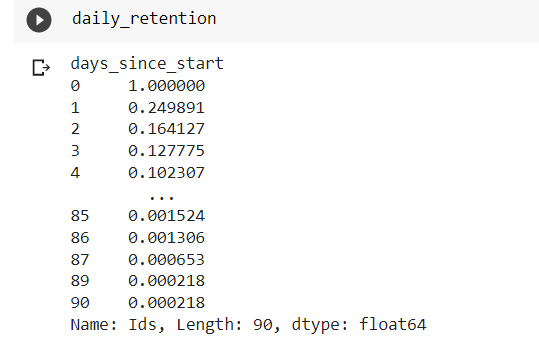
All users stay 1 day since start, while 0.75 stay 2 days since their login, 0.5 stay 3 days since their login.

**Daily Retention = Number of users for a given number of days since start of user/Total unique Users**

Percentage of Daily Retention vs Number of Days since start

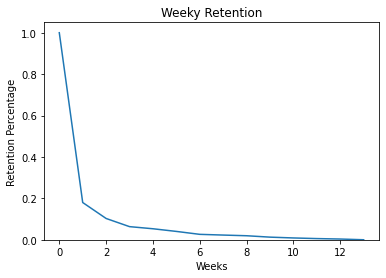


Percentage of Daily retention for days for all users

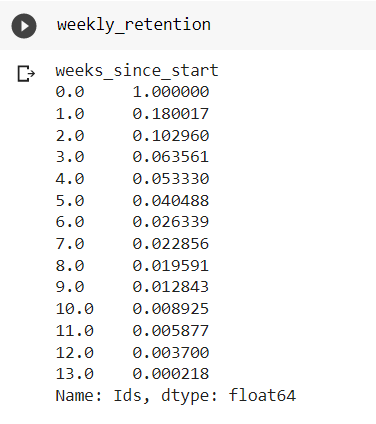


**Weekly Retention = Number of users for a given number of weeks since start of user/Total unique Users**

Percentage of Weekly Retention vs Number of Weeks since start

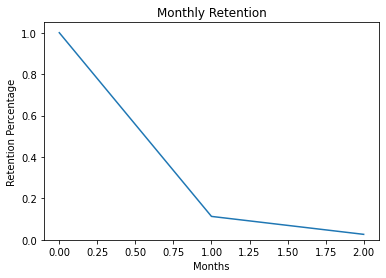


Percentage of Weekly retention for weeks for all users

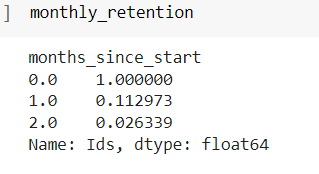


**Monthly Retention = Number of users for a given number of Months since start of user/Total unique Users**

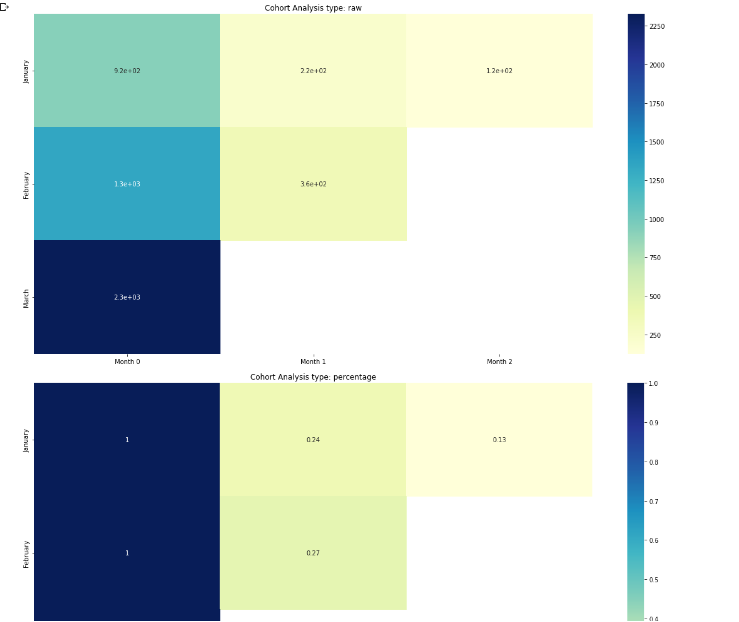
Percentage of Monthly Retention vs Number of Months since start



Percentage of Monthly retention for Months for all users



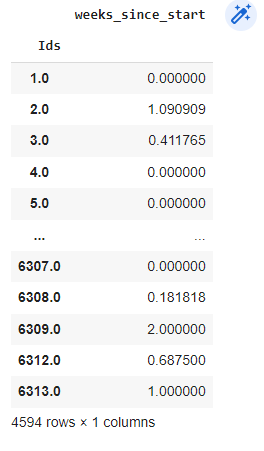
Monthly Retention using Calculate\_Retention Library for raw and percentage



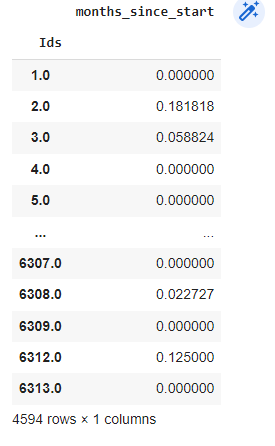
Q4. Day-on-Day Retention Analysis for each User

**Daily Retention (Average Daily Gap) for each user**

****

**Weekly Retention (Average Weekly Gap) for each User**

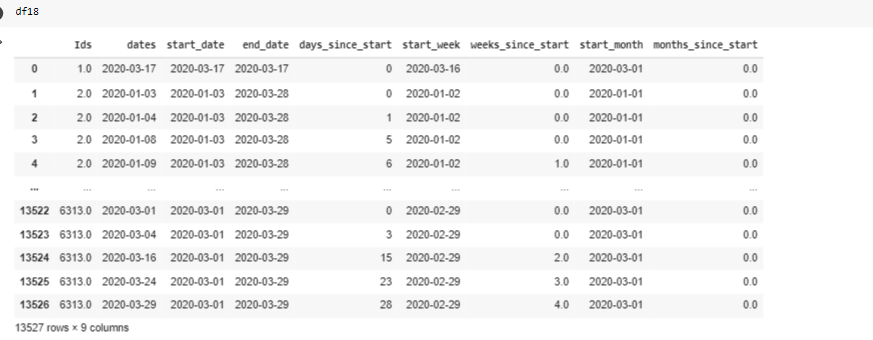
**Monthly Retention (Average Monthly Gap) for Each User**

****

Other Metrics calculated are

Customer Lifetime Value (CLV) = (Total number of User Logins/Number of Unique Users) X Average LifeSpan of User = 18.625 days

Loyal Customer Rate = Repeated Users/Total Users = 0.437



Q5. User Segments and Basis for Segmentation

X-Factors :

1. Login Count

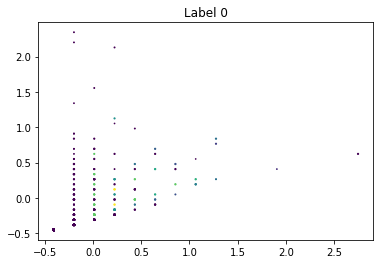
2. Days from Last Login (days)

3. Days between first and last login (days)

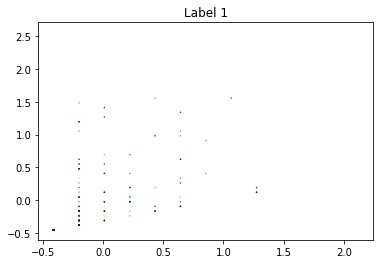
4. Average Gap between logins (in sec)

KMeans Clustering to cluster into 3 Clusters

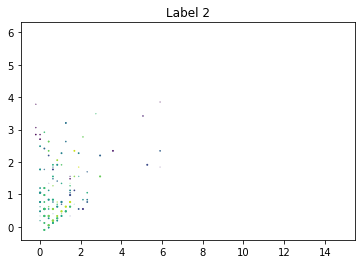
Data Points with Label 0



Data Points with Label 1

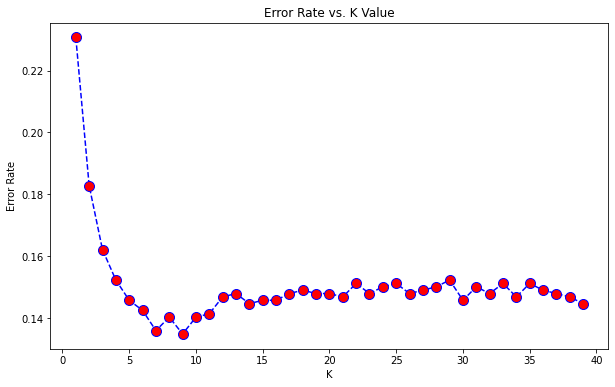


Data Points with Label 2



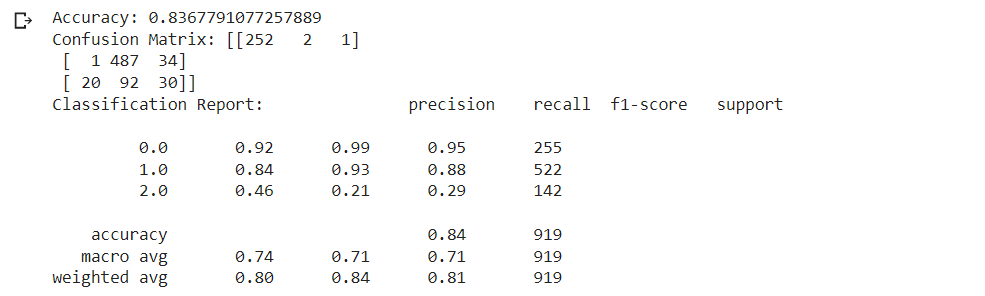
1. KNN with KMeans Clustering

Elbow Curve



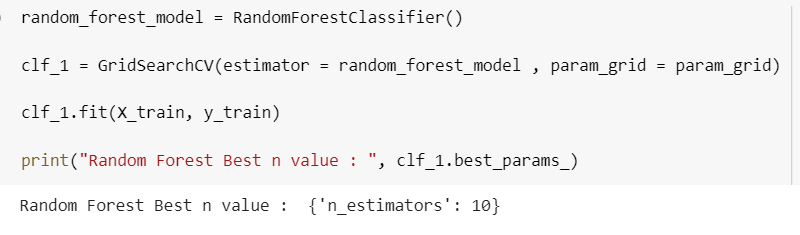
Minimum Error at K=9

Metrics for K=9



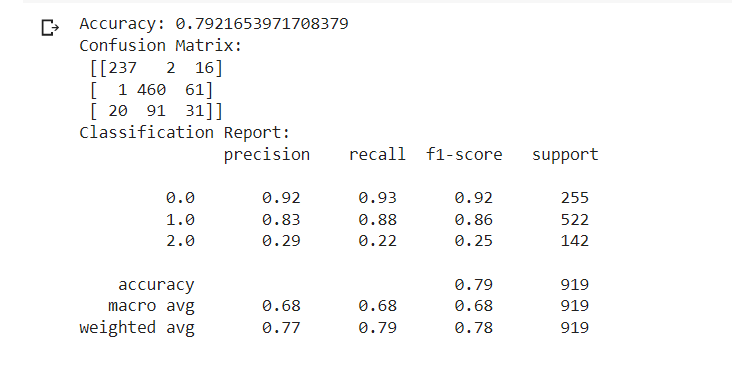
1. RFDT with KMeans Clustering

Use Grid Search to find best number of Decision Trees



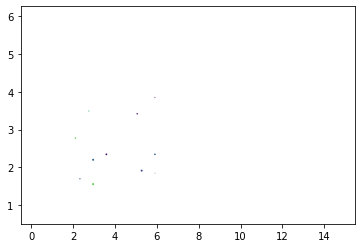
Best for n\_estimators = 10

Metrics for n\_estimators = 10

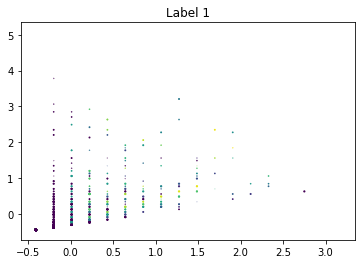


1. Logistic Regression with Birch Clustering

Data Points with Label 0



Data Points with Label 1



While clustering using Birch Clustering, the number of data points in each cluster varies hugely with Label 1 having a lot more data points than Label 0.

Metrics of Logistic Regression

