Marketing Analytics Internship

Andus Kong

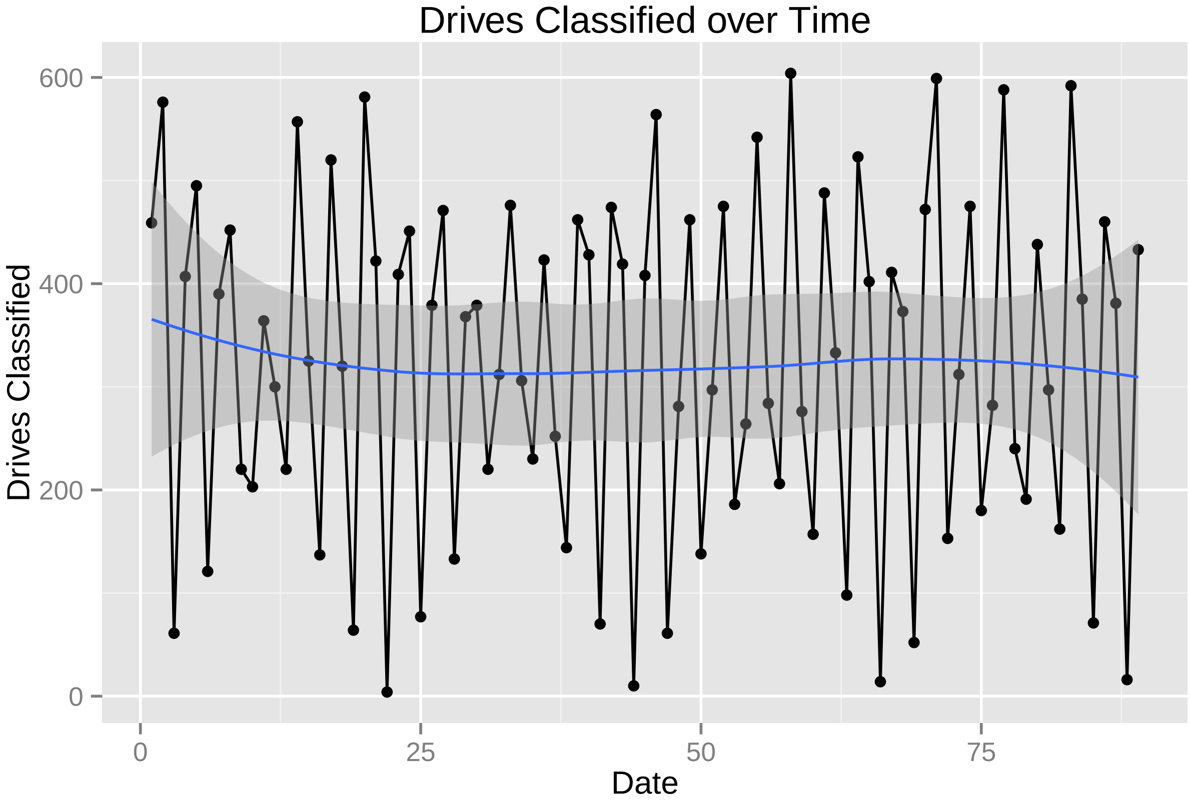
Thursday, July 02, 2015

Goal

Create a behavioral-based email policy to remind users to use MileIQ when they become less engaged.

Exploratory Analysis

I decided to graph each variable over time within different time frames to see if there were any obvious relationships between every variable and time. I used time frames such as all 5000 data points (14 years), one year frames, three month frames, and one month frames.



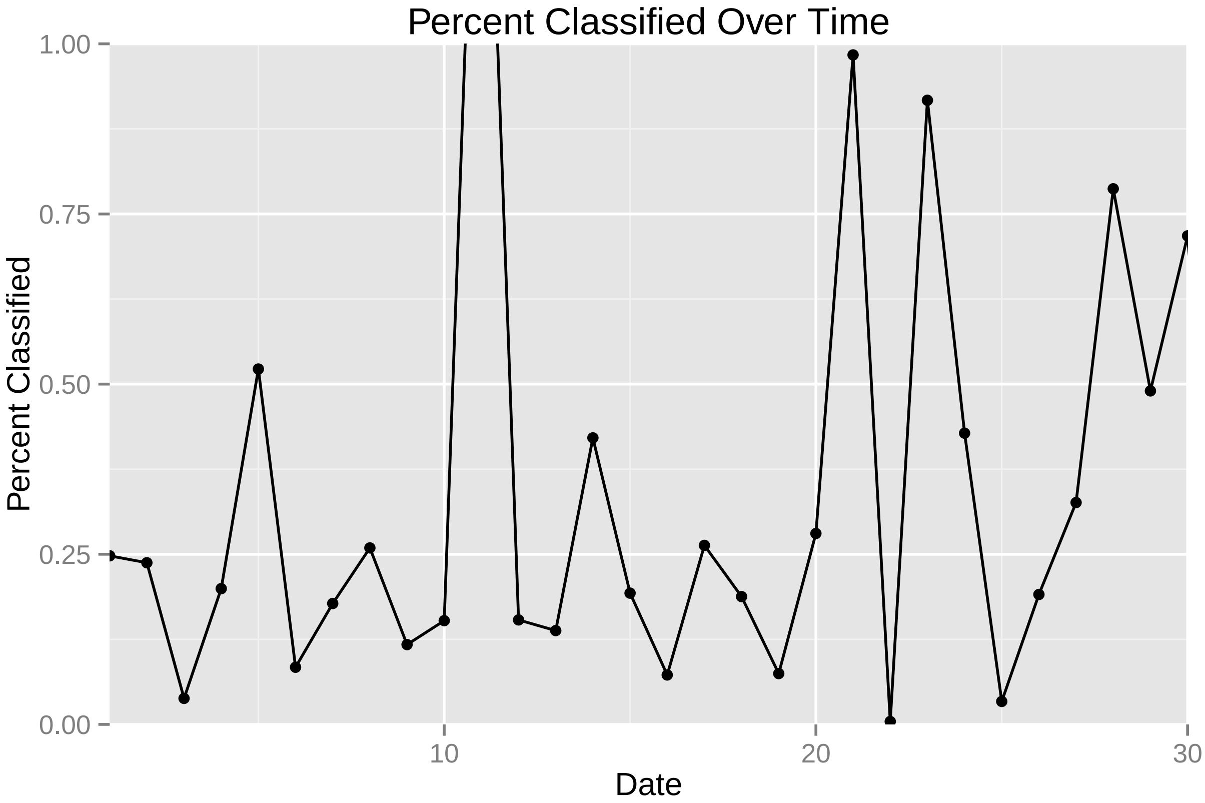
Above you see one example of the first three months of Drives Classified over time. You can see that possibly there seems to be a dip of drives classified within the first 10 days, but when you graph all the one month periods, there seems to be no correlation between Drives Classified and Time. In fact the relationship seems completely random. This holds true for every other variable over time.

Creating Variables

Since there seems to be no obvious correlation, I decided to create new variables and try to find relationships that otherwise weren't apparent.

1. Percent Classified

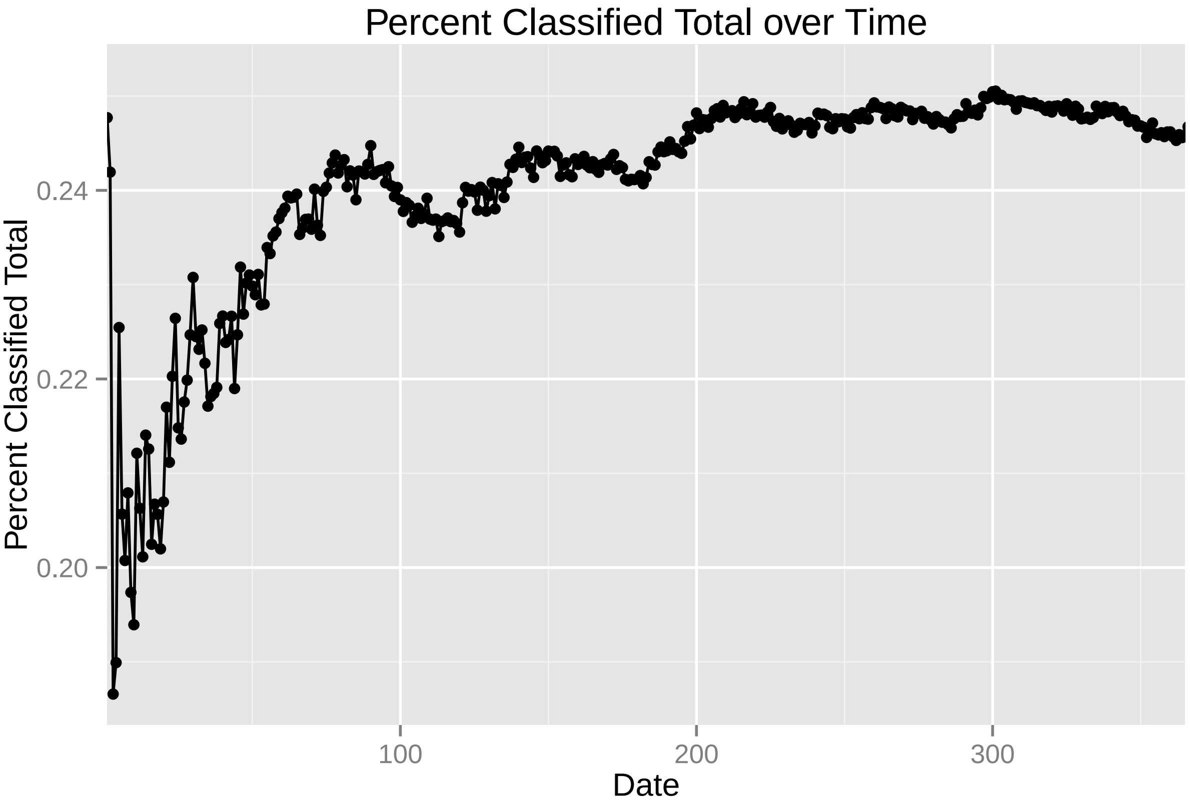
The variable Percent Classified should give us the percentage of Drives Captured that were classified within each day. This can be calculated by Drives.Classified / Drives.Captured. This metric can hopefully give us a measure of how engaged the users are because classifying is using MileIQ to the fullest.



Above you see the graph of Percent Classified over Time for one month. The most startling aspect of this graph is that there are values greater than 100%, which means that more drives were classified than drives captured on some dates. This can be explained by the convenience of MileIQ, which allows users to classify drives days after drives were actually captured. Again when you take many different time intervals, there seems to be no pattern.

2. Percent Classified Total

Since there seems to be no relationship in the Percent Classified, I decided to aggregate the values and created Percent Classified Total, which is calculated by adding the second day's Drives Classified and Drives Captured and taking them over each other. You repeat this process adding an additional day each time. This variable should get rid of the day to day fluctuations we see from the Percent Classified and hopefully we find a relationship.



The above graph illustrates Percentage Classified Total over Time for one year. Sadly, aggregating the percentage classified does not help with the marketing plan, but is still useful. In this scenario, imagine date 1 represents the first day MileIQ is released. We see an initial fluctuation but as the userbase grows we see that the total percentage classified over time levels off at around 25%. That means that in it's thirteen year existance, MileIQ sees its userbase classify 25% of all drives captured. This may be a metric of overall perfomance of MileIQ.

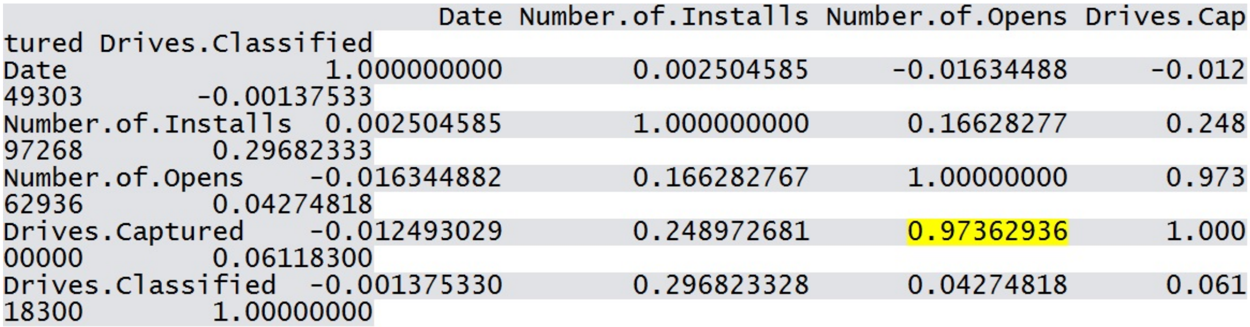
3. Average Classified

Because I identified most engagement as classifying drives, I decided to average day to day Drives Classified and graph that. That is calculated by summing the first Drives Classified with the second and dividing by two, the second with the third, etc... I did not include the graph because it was essentially the same as the Drives Classified graph and there were no relationship no matter what time frame I did.

4. Percent Open

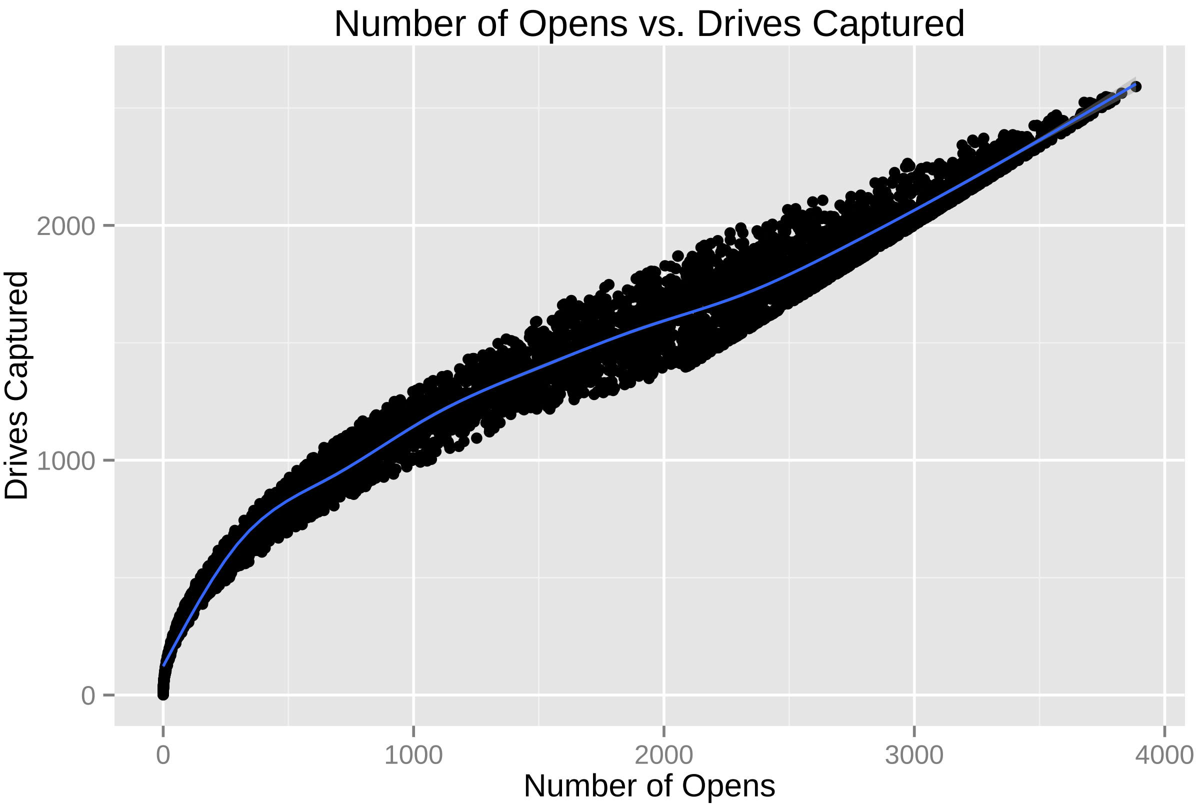
I wanted to calculate the percentage of drives classified given the amount of opens within a day, which is calculated by Drives Classified / Number.of.Opens. By doing this I was hoping to find where usage of MileIQ was high and low; however, there were no relationships of Percent Open over Time. Again I left out the graph.

Correlation Matrix



```
                             Date Number.of.Installs Number.of.Opens Drives.Cap
tured Drives.Classified
Date                  1.000000000        0.002504585     -0.01634488     -0.012
49303       -0.00137533
Number.of.Installs  0.002504585        1.000000000      0.16628277      0.248
97268        0.29682333
Number.of.Opens    -0.016344882        0.166282767      1.00000000      0.973
62936        0.04274818
Drives.Captured    -0.012493029        0.248972681      0.97362936      1.000
00000        0.06118300
Drives.Classified  -0.001375330        0.296823328      0.04274818      0.061
18300        1.00000000
```

Since there seems to be no correlation over time with the variables, I decided to see if there were relationships among variables. According to the correlation matrix above, number of opens is highly positively correlated with drives captured (0.97).



As suspected there seems to be a extremely high postive relationship between number of opens and drives captured, which suggests that a strong predictor of usage of MileIQ is the number of opens. We know that as more users open the app, more drives are recorded. Even though not all drives are classified, there is still engagement by the user.

Conclusion: Behavioral-Based Email Policy

Because of the high correlation between Number of Opens and Drives Captured and subsequent confirmation of the statistical significance of their relationship, I recommend that MileIQ employ a user by user based email system. MileIQ has the data on each individual user, and you should send an email to users if they are not opening the app for three consecutive days. This email should be followed up by another email if the user fails to open MileIQ for an additional three consectuve days. I determined the optimal amount of days by taking each individual one month period and calculating the largest number of opens (local maximum) within that period. I then find the lowest value (local minimum) after the day of the local maximum and calculated the amount of days between those two values. I averaged these values and found that the average length of time between the local maximum and local minimum is 7.37 days. Basically the average length of time within a month between the most user engagement on MileIQ to the least was 7.37 days. To offset the decline of usage, I recommend MileIQ email users after three days of not opening the app to make sure user engagment always stays high.

Other Comments

My initial reaction looking at the data was that its difficult to determine user behavior when you aggregate all the users together. I believe that an even a more accurate email policy can be determined if there was data on each individual's behavior instead of lumping them together. That way you can really see trends and find user behavior.