**Introduction**

With the increase in free-WiFi market, which is now in its peak, free WiFi has become a profitable and one of the vital means for businesses to promote their products and services, with the use of their advertisements, hence increase their brand awareness. Moreover, one of the greatest advantages of offering free WiFi connection at particular locations, companies can collect huge amount about their current and potential clients by relatively saving their advertising budget and target their customers more precisely.Whenever companies offer free WiFi at a specific locations, companies know that they offer a highly demanded service that most of the customers benefit from.

The paper is based on the analysis of WiMedia’s data, it is an A to Z online WiFi advertising platform, the goal of which is to create a global, trusted, effective advertising platform based on public WiFi hotspots and in-depth proximity marketing strategies.Thus, the company has created the platform which is planned to be used by businesses to promote their products and services.

The paper aimed to solve following business problem:

To identify the most appropriate advertisements to show to specific types of people while they try to connect to WiFi.

The analysis are meant to find the ‘win-win’ solution both for the customers and business. By showing the right advertisements to people who will benefit from it the most, businesses will promote their sales and potential customers will satisfy their needs.

**Dataset and feature**

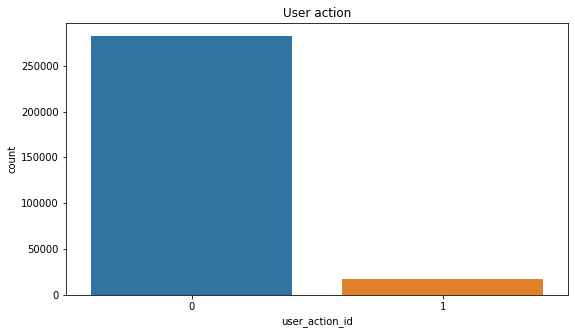
Wimedia’s initial data consisted of more than 1 million rows and total of 11 columns. After doing random sampling, the data has 299,999 rows and 11 columns. Each of the columns provides the following information:

* **Ad\_instance\_id**- is the id is of particular advertisement
* **Segment\_id\_x**- is the segments of the advertisements, examples of ad segments are Finance, Betting, Food products, Mobile Applications etc.
* **Place\_id** - ids for a particular location
* **Segment\_id\_y**- id from a separate table defining the place segments. Examples can be bars, cafes, restaurants, shopping centers, public areas, educational epaces, etc.
* **User\_action\_id**- identifies the action of the users when the ads are shwn.It is null when there is no action from user, is -1 when user skips the ad, is 1 when user clicks on some area of the ad**.**
* **Action\_time**
* **Screen\_height**
* **Screen\_width**
* **User\_ID**
* **mac\_address** - shows the network interface controller in communications within a network segment.

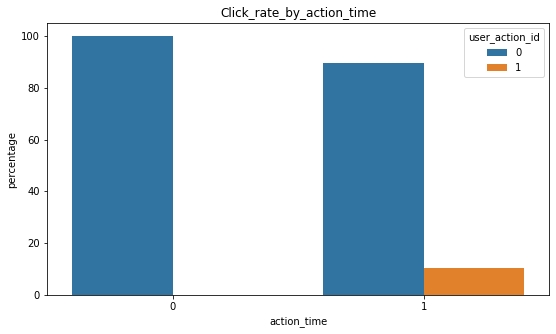
There are 94681 duplicates in the data. However, they are not duplicates in their meaning, because a person, who is shown several ads with a relatively short interval may be included in the same action time group (“0” or “1”), and so are represented as identical rows in our data. Because we used a pivot tables, they didn’t cause any problems (there we have unique rows).

The data does not contain any missing values.

One of the main transformations on the data was to transfer user\_action\_id into two categories instead of three. If user clicks on some area of the ad the variable gets the value of 1, any other action equals to 0. Furthermore, we found the distribution of the user\_action\_id, which was 94.118314% and 5.881686% for 0 and 1 categories, correspondingly.

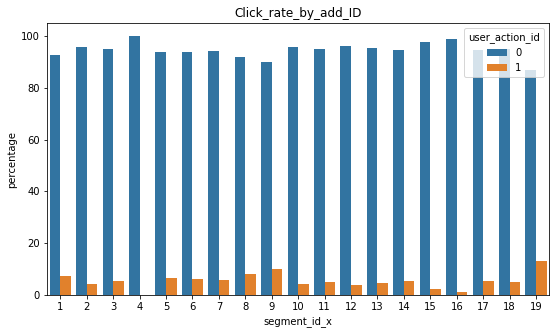


We also did some visualization to understand the relationship between our variables.



For example, from above mentioned graph we see that all the people who have an action time equal to 0 didn’t click at all, and from people who have action time equal to 1 approximately 85% clicked and 15% didn’t clicked.

The next figure shows the percentage of clicking-skipping for all segments of advertisements.



**Methodological approach:**

Two approaches were used to solve the company's problem.

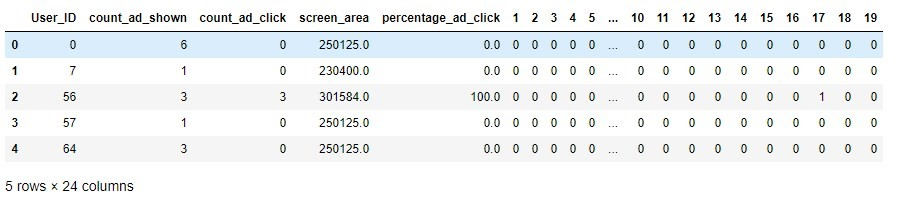
1. Determine whether a user belongs to a certain ad segment, i.e. will he click to the ads of particular ad segment
2. To see whether the segments of ads are constructed in an appropriate manner
3. Determine whether a user belongs to a certain ad segment, i.e. will he click to the ads of particular ad segment

 In an attempt to generate more significant variables following transformations were done:

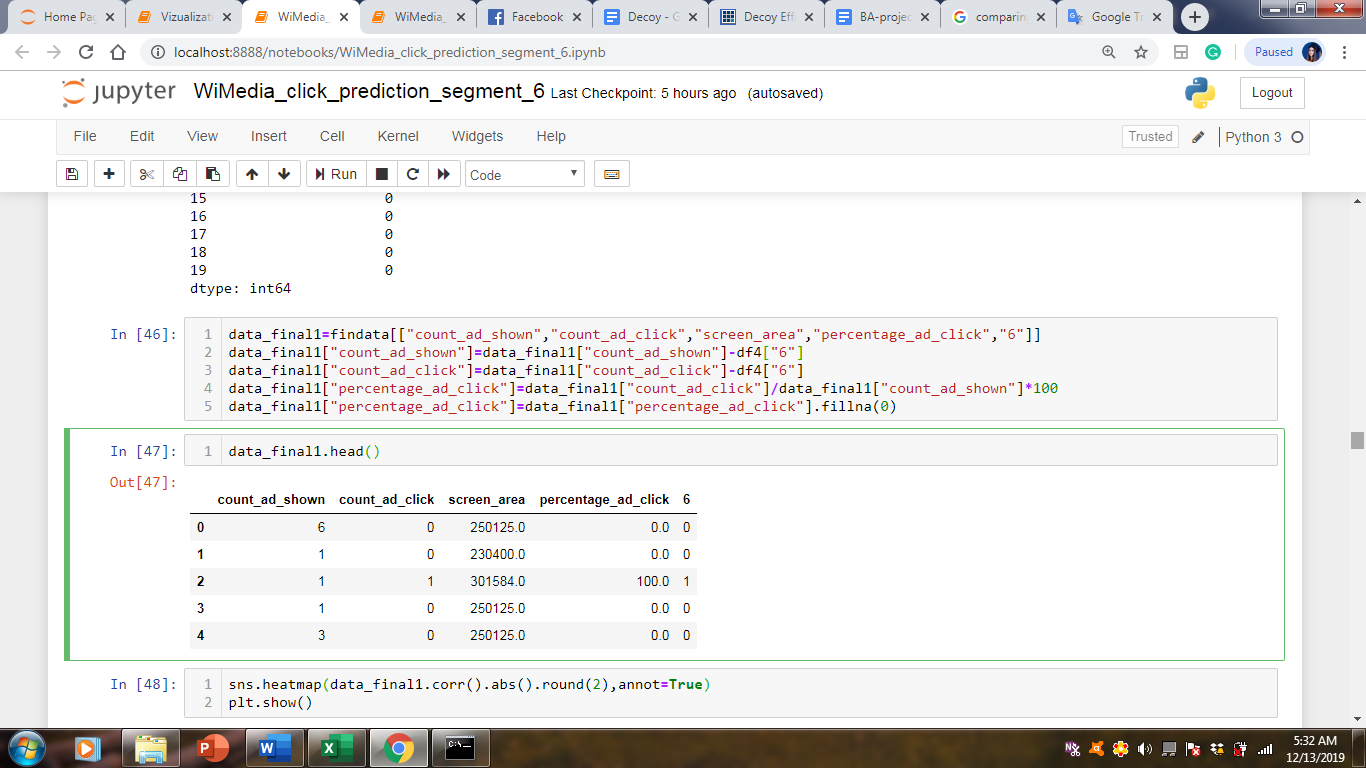
* Screen\_Area- by multiplying screen height and width we computed screen area to decide whether the device’s screen size has an impact on users behavior. Which indeed has (using decision tree classification model it is revealed to be one of the most important features)
* Rotate- by deducting screen width from screen height we found out whether the users hold their phones vertically or rotated.

The next step was to build pivot tables to get the total number of ads shown to unique users, total number of clicks by users to any ads, means of users’ devices’ screen area and total number of clicks to the ads of certain segments. The latter variable was converted to new variable which aimed to state whether the user belongs to a certain ad segment or not; 1 value means that the user belongs to a segment on condition that the number of clicks to that segment ads are at least 1, and 0 if there are no clicks to the segment ads by a user.

  Above mentioned pivot tables were converted to data frames and after several transformations were joined to have final data on which the models were built.

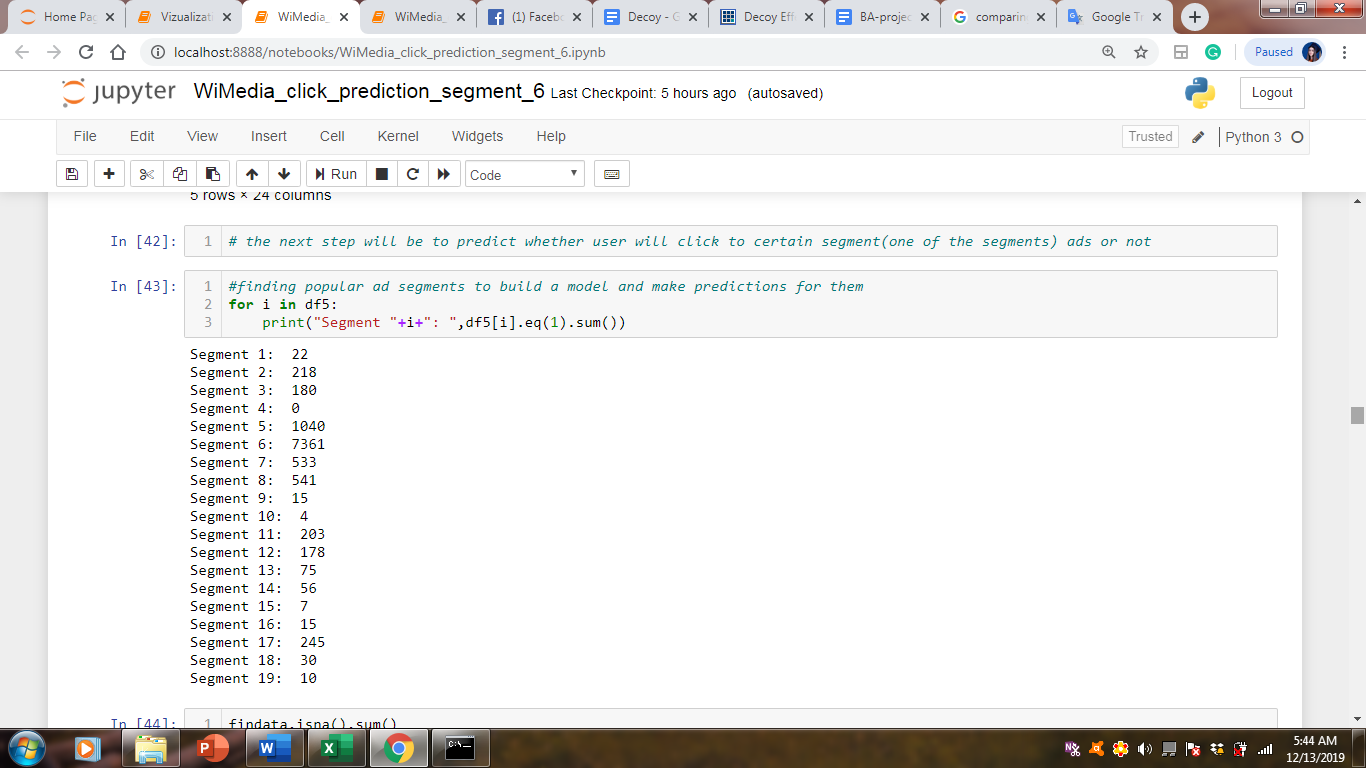


Initial data frame



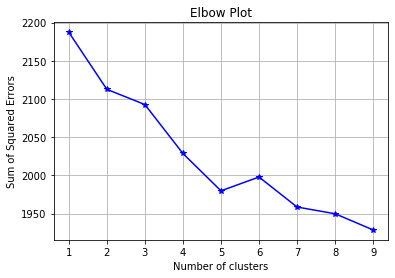
Example of data frame on which the models were built

We have built several models on above shown data frame, where our independent variables are count\_ad\_shown (ads shown to a unique user), count\_ad\_click(total number of clicks by a user to any ads), screen area(means of users’ devices’ screen area), percentage\_ad\_click(number of clicks divided by the total number of ads shown), and our dependent variable is “6”, which shows whether the user belongs to that particular (6) segment or not. By the way, this is our most popular segment.



**Clustering**

In order to specify the optimal number of segments of users, and compare the resulted segments with WiMedia’s generated user segments, K-Means Clustering algorithm was used. To obtain the optimal number of clusters (user segments) the Elbow method has been used. The method suggests that whenever K (number of clusters) increases, sum of squared error (SSE) should decrease. However, K-Means Clustering failed to work on the data as when number of clusters increases up to K=5, SSE decreases gradually, but when number of clusters equal to 6 SSE significantly increases. This can be explained by the number of 0s in the dataset, which are incomparable more than 1s.



**Results**

The built models on the data:  Logit, Decision Tree and Random Forest tuned with hyper parameters showed the following results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Segment\_6 | Segment\_5 | Segment\_7 | Segment\_8 |
| Logit |  |  |  |  |
| ROC AUC train | 0.51 | 0.62 | 0.66 | 0.56 |
| ROC AUC test | 0.52 | 0.64 | 0.67 | 0.6 |
| Mean 3-fold ROC AUC | 0.51 | 0.63 | 0.66 | 0.59 |
| DT |  |  |  |  |
| ROC AUC train | 0.77 | 0.8 | 0.77 | 0.78 |
| ROC AUC test | 0.76 | 0.8 | 0.76 | 0.78 |
| Mean 3-fold ROC AUC | 0.76 | 0.79 | 0.75 | 0.77 |
| Random Forest |  |  |  |  |
| ROC AUC train | 0.77 | 0.81 | 0.81 | 0.78 |
| ROC AUC test | 0.77 | 0.79 | 0.77 | 0.78 |
| Mean 3-fold ROC AUC | 0.77 | 0.8 | 0.78 | 0.79 |

 For all 4 cases Random Forest works best, as its AUC ROC is the highest for moth Train and Test datas (only for segment 5 ROC AUC test of DT is a bit higher than RF).

However the second approach , which was done using K-Means Clustering, failed to work on the data as when number of clusters increases up to K=5, SSE decreases gradually, but when number of clusters equal to 6 SSE significantly increases. This can be explained by the number of 0s in the dataset, which are incomparable more than 1s.

**Recommendation and future plans**

In future, before showing particular add to particular user, WiMedia could use one of our models, particularly Random Forest, which is the best one, to predict whether he or she will click to the particular add with approximately 77-80% precision level, and then decide to show a particular add or not.

**References**

* <https://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf> ,
* <https://www.tanaza.com/tanazaclassic/wifi-as-an-advertising-tool/>,
* https://lyst.github.io/lightfm/docs/lightfm.html
* <https://www.academia.edu/41230478/Cluster_Analysis_of_Customer_Reviews_Summarizing_Customer_Reviews_to_Help_Manufacturers_Identify_Customer_Satisfaction_Level?fbclid=IwAR009PVVMfKStZv31Rpjqbhqttt4m3Eq9TU02PeSn8WqYjfmTBpYvWtJkUA>