***INSAID TELECOM - Customer First***



Capstone 1 - EDA - Submitted by Team 1008

Table of Contents

[**Introduction**](#_moaidao0auq6) **2**

[**Project Description**](#_ah62y47mjo2v) **2**

[**Problem Statement**](#_m7izm589ree5) **3**

[**Problem Analysis**](#_9sh6ndx55rye) **3**

[**Sources of Data**](#_n4hi4019ohrk) **4**

[**Summary of Data Mining**](#_yacdhjkql563) **6**

[**Proposed Solution for the Client**](#_90rdd3ytgi) **9**

[Insights from the Analysis](#_88soygcgwkck) 9

[Conclusion](#_sfmjuf3t22qc) 12

[Recommendations](#_mwcwr7s4bqkj) 12

[**Appendix**](#_a5zi6c3jz54e) **12**

[Tools](#_qouik4673dsj) 12

[DS Tools](#_hu658wo5eyoc) 12

# 

# Introduction

As a customer, it is always wonderful to see services customized to your needs. Businesses try to understand your behavior and adjust their offerings to ensure that you feel the need to continue to stay with their services. In the highly competitive telecom industry, customer service and customer experience are critical factors in building and maintaining a competitive advantage.

**InsaidTelecom**, (here on referred to as ‘client’), one of the leading telecom players in the country, understands and believes that customizing products to serve users needs, is very important for its business to stay competitive.

Currently, they are seeking to leverage behavioral data from more than 60% of the 50 million mobile devices active daily in India to help its clients better understand and interact with their audiences.

# Project Description

Team 1008 was tasked to analyze the data collected by the client and push the analysis to the team building the dashboard. This dashboard would be used to understand the user's demographic characteristics, based on their mobile usage, geolocation, and mobile device properties.

The Data was collected from mobile apps that use the services provided by the client. Full recognition and consent from the individual users of those apps have been obtained, and appropriate anonymization has been performed to protect privacy. Due to confidentiality, details on how the gender and age data was obtained, have not been provided, however their accuracy has been accepted as the ground truth, as instructed.

Doing so will help millions of developers and brand advertisers around the world pursue data-driven marketing efforts which are relevant to their users and catered to their preferences.

# 

# Problem Statement

The client has collected vast amounts of customer data, and would like to leverage the same in order to provide a better user experience to their customers. The challenge is to extract actionable insights to drive decision making towards the goal of customer satisfaction, and growing the business to increase market share.

To this end, the following pertinent questions can be asked:

* What is the typical profile of the client’s user base?
* What are the typical usage characteristics?
* What are the types of devices that the users favour?
* What can the client do to attract and retain more customers?

# Problem Analysis

In order to address the user demographics questions, the available data needs to be gleaned through to understand the kind of data that is available. Is this data / information enough or will the client need to improve / change their methodology entirely?

The dataset provided was 3 parts:

1. User data consisting of the gender, age, age group and the associated device id
2. Device information consisting of phone brand name, model type and the associated device id
3. Events data consisting of an event id, location (lat/long), timestamp and the corresponding device id. This was recorded each time the device logged into the client’s network to avail of a service

The first concern was the mapping of the three sources of data with each other. In this regard, the device id being the common variable was selected as the anchor, based on which the mapping of the data points would be undertaken.

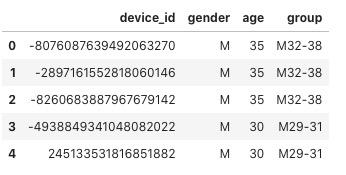
Once this mapping was done, the next task would be to look for any missing information etc., and then analyse the merged dataset to assess whether the information obtained would be adequate to pursue our stated goal of gathering actionable insights.

# 

# Sources of Data

User data and device information was provided by allowing us to connect to the client’s servers and extracting the information using MySQL. The MySQL Connector driver for Python was used to connect to the servers through a Jupyter Notebook and the downloaded data was saved in csv format.

The user dataset consisted of over 74 thousand data points, having 4 variables. These are the device id, gender of the user, age and the relevant age group.

An example is shown here:

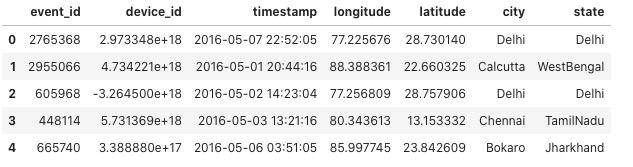
The device dataset consisted of over 87 thousand data points, having 3 variables. These are the device id, device brand and the device model.

An example is shown here:

Note that the phone brand and device models have some of the data in simplified chinese.

The events dataset was sent via link to a direct download from the client’s servers. The dataset consisted of over 32.5 lakh data points, having 7 variables. These are the device id, event id, time stamp, location data in the form of latitude and longitude, city and state.

An example is shown here:



Since the device id is the common variable in all three data sets, this was used as the anchor for mapping the data points and merging / syncing the data sets for further analysis.

Additionally, the client has requested that we focus our analysis for the state of Delhi only. The final dataset for analysis, filtered for the state of Delhi after merging of the datasets, contains just over 7.5 lakh records, each record consisting of 12 variables.

An example of the data is here below:



Note that the above image shows the data after cleanup and final merging.

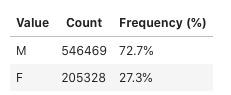
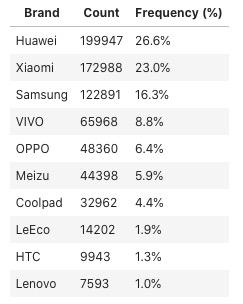
# Summary of Data Mining

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Observation** | | | **Resolution** |
| **Observation Base** | **General Observations** | **Issues / Potential Issues** |
| 1 | Events database | * 3252950 unique event ids and 60865 unique device ids * events data captured from 2016-04-30 23:52:24 to 2016-05-08 00:00:08 * more number of events are seen in the month of May * more number of events(43) seen on 2016-05-03 10:00:03 * 32 unique states and 933 unique cities * Event data has no duplicate values | * state -> missing 377 values * longitude -> missing 423 values * latitude -> missing 423 values * device\_id -> missing 453 values * There are multiple devices having similar lat (41.87 N) and long (12.56 E) tracing to different states Rajasthan, Tamil Nadu, Delhi etc whereas these coordinates (41.87N, 12.56E ) actually are of Rome (Italy) * -1448078833416777984 (present in 1368 rows in whole data and three times in the delhi filtered data) device id traced in Delhi having coordinates (41.87N, 12.56E) on 1st May 2016, 4th May 2016, 5th May 2016 while same device is logged in coordinates of Delhi (28 N, 77 E) on 1st May 2016, 4th May 2016, 5th May 2016. * Locations plotted in Rome, Kabul and Dubai | The logic followed for data cleanup is as follows:   * Rows with missing state values had the city value   therefore the missing state values were mapped to the the corresponding values from the same city in other rows   * Lat / Long pairs were unique to the device\_id   the missing lat / long values were mapped to the corresponding values from the same device\_id in other rows. Although some device\_id had multiple lat / long pairs at various event\_ids, the ones with missing lat /long had only single pairs   * Since Lat / Long pairs were unique to the device\_id we used this to our advantage to fill in missing device\_id info. The rows with missing device\_id had all other info, therefore we mapped the missing device\_id to the corresponding one with the same latitude value * For the data points where a device\_id has multiple locations, we have left them in place. They will either be treated as 'roaming' events or they will be corrected as location errors depending on the analysis * This will be determined after plotting the final cleaned and merged database on a map. |
| 2 | User Database | * 74645 unique device\_id * Male users are about 47904(64.17%) of 74645 * Minimum age - 1 year Average age - 32 years Maximum age - 96 years * Male users of the age group 23-26 are more than any other group | * Clean dataframe. no intervention necessary | * Clean dataframe. no intervention necessary |
| 3 | Device Database | * 87726 unique device ids * 116 unique phone brands * 1467 unique device models * 116 unique phone brands * Top used phone brand - Xiaomi * most used model - Redmi Note | * phone\_brand and device\_model have Simplified Chinese characters | * Phone brand and Device model names in chinese were mapped to English. This was done using Google Translate API to get the required translation and make the conversions * extracted list of unique phone\_brand -> sent list to a dataframe * used multiple available translation modules to translate each item in list from simplified chinese to english -> captured these into separate dictionary pairs * merged these into a list of dicts and converted the list into a dataframe for scrutiny * manually check the dataframe rows for unusual / incorrect / garbage translation and manually create new dict with the most correct / likely translation (easy process as only 116 unique items) * use the new dict to replace the brand name from chinese to english * Same process was used for phone model |
| 4 | Merged Database |  | * device\_id in float format in events data and in int64 format in user and device data * this causes major issues when merging dataframes, with most of the info from user and device dataframes being lost | * device\_id used as the key to merge the three databases. * First convert data type for device\_id as float in all dataframes * merged the user dataset to device dataset * The merged dataset was then merged with the events dataset. This preserves maximum data |
| 5 | Delhi Database | * The final dataset was filtered to only contain data for Delhi | * No issues were found |  |
| 6 | plotting delhi data onto a map | * plot of the locations revealed a very interesting scenario * vast majority of the points are geo-restricted to a quadrangle limited to north east delhi * remaining points are scattered around delhi * Some devices registered in delhi are being used in the western part of the neighbouring state of UP | * Since each device\_id has only one geolocation, any event is therefore only going to plot to that one location. This will cause multiple event points to be plotted one over the other which is a waste and also computationally time consuming. (only delhi data took 25 mins and consumed 16GB ram) | In order to reduce compute time for plotting:   * an event count was taken with data grouped by device\_id, long , lat and sent to a dataframe. another dataframe with duplicates dropped based on device\_id was then filtered out. these two were then merged to get a new dataframe with single row of data per device \_id with all prior info but with a new column containing the counts of events for that device\_id * This yielded 4912 data points to plot VS 751797 points. * compute and plotting time reduced to 10 seconds! * plotting accomplished using folium module that uses OpenMaps |

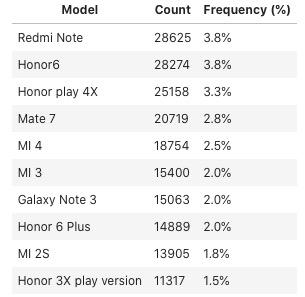
# Proposed Solution for the Client

## Insights from the Analysis

Adding observations from my analysis to help you complete the compilation faster

* Final Delhi Dataframe has 751797 records.
* There are 4909 unique devices
* 61 unique brands
* 631 unique device models
* Data spread for the week shows:
  + the ramp up in usage from the weekend towards mid week
  + the ramp down in usage towards the weekend
  + peak usage is on Tuesday
* Data spread by time of the day indicates:
  + peak usage is after the start of work hours, about 10 am
  + there is a drop down in usage towards mid afternoon
  + again an increase in activity towards 8 pm
  + a minimum of ~30% activity is recorded even at the dead of night
* Male Female Ratio
  + 
* Age wise distribution of users reveals the following:
  + min age is 16 years
  + max age is 94 years
  + mean age of the user is 34 years
  + median age is 32 years
  + Interquartile range is 13 years
    - IRQ between 27yrs and 40 years
* Usage trends by hour of the day / day of the week
  + usage trends show that activity picks up from 6 am and peaks mid morning on all days
  + usage then drops todays the latter part of the day before a shorter peak at 8 pm
  + activity continues to drop thereafter
  + activity is lowest between 3 and 4 am on all days
  + Wednesday has the highest activity peak for any day, at 10 am
  + Tuesdays has an unusual peak at 7 pm compared to all the other days
* Usage trends in males by hour of the day / day of the week
  + usage activity for the week plotted by the age groups in males shows the following:
    - the age group 32-38 and 39+ are the most active groups
      * their peaks in activity are at 10 am and 11 am respectively
      * the difference in activity level compared to other groups is likely from income generating activity
    - the 29-31 and 23-26 age groups are the next most active
    - the 27-28 group usage is unusually lower than their older and younger groups
    - 22 group is the least active
* Usage trends in females by hour of the day / day of the week
  + usage activity for the week plotted by the age groups in females shows the following:
    - the 33-42 and 43+ groups are the most active
      * peak activity is at 7am and 6am respectively
      * indicates earlier start to the day as compared to males in the earlier graph
    - all other groups usage is generally closely trending, however the 20-32 age group has the highest usage among the remaining
* The TOP 10 Brands in use:
  + 

# 

* TOP 10 Device Models
  + 
* Preference of the Brands by Gender:
  + Males and Females have the same preference of the top 4 brands
    - Huawei, Xaomi, Samsung and Vivo
  + for place 5
    - Males prefer Meizu
    - Females prefer Oppo
  + LeEco, HTC and Lenovo have NO USERS in the Female gender

## 

## Conclusion

## Recommendations

# Appendix

## Tools

### DS Tools