**Consulting Report**

Contents

[1. Introduction 2](#_Toc98095308)

[2. Project Description 2](#_Toc98095309)

[3. Problem Statement: 2](#_Toc98095310)

[4. Problem Analysis: 2](#_Toc98095311)

[5. Sources of Data 2](#_Toc98095312)

[6. Summary of Data Mining (What kind of challenges you faced with the Data and how you resolved them? Summary of your Analysis): 3](#_Toc98095313)

[7. Proposed Solution for Customers (Describe your Analysis in Detail) 4](#_Toc98095314)

[8. Tools 5](#_Toc98095315)

[9. Conclusion 5](#_Toc98095316)

## Introduction

*This project is to work on a Telecom service provider data to find out mobile and application usage by different mobile brand, model, gender, age group of users etc.*

*Based on the usage and analysis report provided by the project team the telecom company will try to follow the conclusion to increase productivity*

## Project Description

**InsaidTelecom**, one of the leading telecom players, understands that customizing offering is very important for its business to stay competitive.   
Currently, InsaidTelecom is 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.

## Problem Statement:

In this **consulting assignment**, Insaidians are expected to build a dashboard to understand user's demographic characteristics based on their mobile usage, geolocation, and mobile device properties.   
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 Analysis:

To help the customer the project team is expected to have depth of clarity in the underlying data.

How much effort need to put into cleansing and purifying the data will decide how closely team is going to look at the data.  
How detailed is the observation stated in the submission report and finally how well the project team presents their consulting journey.

As this is analytics consulting hence, project efforts in terms of finding user behaviour is going to directly impact the company's offerings.   
So, project team need to help the company understand what is the right way forward and suggest actionable insights from marketing and product terms.

## Sources of Data

There are three datasets:

One dataset is in .csv format and other two are extracted from MySQL server host “cpanel.insaid.co”

**events\_data.csv** is downloaded from below URL,

<https://drive.google.com/file/d/1Ir3rW0YTKmk7MSjVjCU_UGMQevhe1v9W/view>

Column names in that dataset are:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| event\_id | Event ID of that time of mobile use |
| device\_id | Unique ID of a particular mobile handset |
| timestamp | Timestamp of event |
| longitude | Longitude of the mobile location |
| latitude | Latitude of the mobile location |
| city | Name of Indian city where mobile is used |
| state | Name of Indian state where mobile is used |

**gender\_age\_train** dataset is downloaded from MySQL server and columns names are:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| device\_id | Unique ID of a particular mobile handset |
| gender | Male or Female |
| age | Age of mobile user |
| group | Age group of mobile users |

**phone\_brand\_device\_model** dataset is downloaded from MySQL server and columns names are:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| device\_id | Unique ID of a particular mobile handset |
| phone\_brand | OEM of mobile handset |
| device\_model | Model name/number of phone |

The three datasets are merged based on primary key = device\_id as that is only the common reference in all the dataset

## Summary of Data Mining (What kind of challenges you faced with the Data and how you resolved them? Summary of your Analysis):

1. There are no missing / null values in the SQL retrieved datasets but there are null values in the events dataset.

Null values are present in the form of device\_id (453), longitude (423), latitude (423) and state (377) in the events\_data.csv file

Solution: We have first used Geopy Python package to fill those null values of state, as the consulting project is specifically for Maharashtra only. By feeding the coordinates into Geopy, we were able to obtain their corresponding state. The other null values were dealt with as described in the subsequent points, as reducing it to Maharashtra removed a lot of the records with null values.

1. The project team had selected the data for Maharashtra and observed that there are 6,77,168 total records for Maharashtra, in which it is observed that:
2. there were 63 null values for longitude and latitude
3. those 63 null values correspond to 3 users (21 records with null values for each)

Solution: We observed that those 3 individual users used their mobiles much more that those 21 times (over 1000 times for each), and where the coordinates were present, each of the 3 users had their own unique coordinate pair. Thus, the null values (longitude and latitude) in the 21 incomplete records for each user were filled with their corresponding unique coordinate pair as taken from their other complete records.

1. There are discrepancies noticed in latitude and longitude - The latitudes and longitudes (9 records specific to Maharashtra state) when plotted using Folium package were showing places outside India, and crucially to the west of Maharashtra

Solution: These 9 records are excluded from the consideration set. All 9 erroneous records were west of 70 degrees longitude, whereas all the other records were east of that value. Based on that condition, the exclusion was carried out.

1. We noticed that in the trimmed down Maharashtra dataset, there are 72 rows with device\_id being null. We also noticed that each device\_id had a unique coordinate, and that the aforementioned 72 rows had 3 unique coordinates (24 rows each).

*Solution:* We observed that those 3 coordinates were also present in other rows (over 1000 times for each coordinate) where the device\_id is present. Thus, the null values (device\_d) in the 24 incomplete records for each coordinate were filled with their corresponding unique device\_id.

1. **gender\_age\_train dataset** There are total 6 unique age groups for both females and males. The group for Males ranges from M22 to M39+ and for Females it ranges from F23 to F43+. The age segments were not the same for males and females, and the size of the segments were different.

*Solution: We created a new age segment field, where each segment covered an interval of 2 years. Since we knew the age of each user, we could generate this field ourselves. We kept the original given age segment data as well.*

1. We noticed that some of the phone brands had Chinese characters.

Solution: Trimming the dataset down to just Maharashtra and then only taking the top 10 records (as per the project restrictions) resulted in a dramatic reduction in the number of records that needed translation. Most of the remaining translation could be done by using the translation matrix provided to us to alter the phone brands database. Only one phone brand (Tianyu) was missing in that translation matrix, so we translated and added it to the translation matrix beforehand so that the database could be fully translated.

## Proposed Solution for Customers (Describe your Analysis in Detail)

1. The two greatest geographical clusters among the users are Mumbai and Pune residents, contributing to 51% and 17.2% of the total Maharashtra user base. The remainder are distributed evenly across Maharashtra, with the various cities having user counts comprising less than 0.9% of the total. As such, any large-scale geographical marketing attempts need to focus on Mumbai and Pune.
2. The most popular brands are Samsung, Xiaomi, Vivo, Huawei, Meizu and Oppo, with the popularity steadily decreasing through that sequence. There are no extraordinarily popular brands, so general brand-based development and marketing will need to cover multiple brands.
3. The user base is skewed in favor of males, with males outnumbering females by slightly more than 2:1. Thus, development and marketing that targets males would have a greater benefit. However, there is still a sizeable female contingent among users, which warrants some effort in that direction as well.
4. The user base covers a total range of 22 to 66 years, but is significantly dominated by young adults (ages 24-32) who comprise 62% of the user base. Therefore, for optimal gains, developers and marketers needs to target that age range.
5. Among males, Samsung is the overwhelming favorite, with twice the user count of the next most popular brand (Huawei). On the other hand, the most popular brands among females are Xiaomi and Vivo, both of whom have equal user counts that are over twice that of other brands.
6. When it comes to gender, the distribution for males and females follows the same trend, as described in point 4. From the perspective of age, there is no value in differentiating between males and females.
7. When it comes to usage times across the day, there is (as expected) a significant bias in favor of day time. The usage statistics has 3 peaks, between 7AM-8AM, 10AM-11AM and 8PM-10PM. There is a slight dip in usage from 8AM-10AM, a slightly larger dip from 11AM to 7PM, and a (expected) huge nighttime dip from 0AM to 6AM. Depending on the development and marketing goals, the strategy may need to be modified accordingly.

## Tools

* DS Tools, Common Python Packages:
  + Numpy – This is a support library for general python use. It provides access to a highly efficient array and matrix processing system, and is a prerequisite for more specialized python modules as well.
  + Pandas – This is a specialized library for easy database creation, management and manipulation, built as an additional layer over numpy. It is the primary tool used in this project for handling the data that is being analyzed.
  + Geopy – This is a python client for accessing online geocoding services. It is used to enhance the analytical potential of the collected data by linking the same to larger geographical survey databases.
* Visualization tools
  + Matplotlib – This is a general visualization tool in python, providing access to a bevy of functions for quick and simple data visualization.
  + Seaborn - This is a specialized library for data visualization built on top of matplotlib, providing more specialized and polished visualization techniques that can seamlessly integrate pandas data structures. This is the primary tool used for data visualization in this consulting project.
  + Folium – This is a specialized library for visualization of geographical data, providing us with the means of creating polished, high-level interactive maps of the collected data.
* RDBMS Connectivity tool
  + mysql.connector – This is a tool for runtime access to mysql databases. This is used in this project to access some of the databases under study.
* ~~Web UI Tools (You don’t have to explain this section)~~
  + ~~PHP~~
  + ~~JavaScript~~

## Conclusion

*Final conclusion to be documented based on final team discussion*