**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 used Geopy Python package to fill those null values of state.

geopy is **a Python package** for developers to locate the coordinates of addresses, cities, countries, and landmarks across the globe

The null values of device\_id, longitude & latitude are solved in step 6. ii.

1. Project team have selected the data for Maharashtra and observed that, there are 6,77,168 total records for Maharashtra in which it is observed:
2. 63 null values for longitude and latitude
3. 3 user’s longitude and latitude are missing (21 records for each)

Solution: Observed that those 3 individual users were using mobile from their single and respective locations so their longitude and latitudes were same. The null values (longitude and latitude) are filled for those users from their available 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 Maharashtra and is located as Kabul

Solution: These 9 records are excluded from the consideration set

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. We observed that the coordinates present in the rows with missing device\_id were also present in other rows.

*Solution: By finding all the unique device\_id + longitude + latitude combinations in events\_data.csv, we could fill in the missing device\_ids using that record's given coordinates. This worked, as the 72 rows with missing device\_ids comprised of three unique coordinate pairs that are also present in hundreds of other rows*

1. **gender\_age\_train dataset** 1. The Age of Males range from 1 to 90 years and for Females it ranges from 10 to 96 years. 2. 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+. There are no NULL values in all the columns.

*Solution: Upon merging the three datasets, we can remove the outliers We can use distribution plots to analyse female and male age groups?*

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

*This section to be filled up from the EDA and what project team will propose to the customer*

## 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*