Consulting Report

INSAID

Batch: 1002

Capstone Project

Table of Contents:

Introduction

Insaid Telecom, 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 behavioural 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

In this assignment, we have studied the demographics of a user (gender and age) based on their app download and usage behaviours. Data is collected from mobile apps that use InsaidTelecom services. Full recognition and consent from an individual user of those apps have been obtained, and appropriate anonymization has been performed to protect privacy. We have treated them as accurate ground truth for the sake of prediction. The data schema can be represented in the following table:

gender_age_train - Devices and their respective user gender, age and age_group.

phone_brand_device_model - Device ids, brand, and models phone_brand: few brands are in Chinese.

Brand English Mapping

'华为'

   'Huawei'

'小米'

   'Xiaomi'

'三星'

   'Samsung'

'魅族'

   'Meizu'

'酷派'

   'Coolpad'

'乐视'

   'LeEco'

'联想 '

   'Lenovo'

events_data - When a user uses mobile on INSAID Telecom network, the event gets logged in this data.

  Each event has an event id, location (lat/long), and the event corresponds to the frequency of mobile usage.

 timestamp: when the user is using the mobile.

Problem Statement

In this consulting assignment, we 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 created to their preferences.

Problem Analysis

We are expected to have a depth of clarity in the underlying data. The amount of effort put into cleansing and purifying the data have affirmed how closely we have looked at this data. This report represents our consulting journey. This being an analytics consulting, our efforts in terms of finding user behaviour will directly impact the company's offerings. There is a pressing need to help the company understand what is the right way forward thereby suggesting actionable insights from marketing and product terms.

Source of Data

InsaidTelecom stores consumer data in the following tables:

1> events_data - when a user uses mobile on INSAID Telecom network, the event gets logged in this data. Each event has an event
2> id, location (lat/long/city/state), and the event corresponds to the frequency of mobile usage. timestamp: when the user is using the mobile.

Description:

event_id-

Unique ID of the instance of usage of Mobile.

device_id-

IMEI of mobile device.

Timestamp-

Timestamp when the user used the device in the format of DD-MON-YYYY HH24:MI: SS

Longitude-

longitude of the location of device usage.

Latitude-

Latitude of the location of device usage.

City-

City of Device usage.

State-

State in which a device is used.

2> gender_age_train - Devices and their respective user gender, age and age_group.

device_id-

IMEI of mobile device.

Gender-

Gender of user.

Age-

Age of user.

age_group-

    Age group of user.

3> phone_brand_device_model - This table stores information about device's brand and models.

device_id-

    IMEI of mobile device.

phone_brand-

    Phone Brand.

device_model.

    Device Model.

Note: that few brands are in Chinese.

Brand Name ---Brand English Mapping

'华为'	'Huawei'
小米	'Xiaomi'
三星	'Samsung'
'vivo'	'vivo'
'OPPO'	'OPPO'
魅族	'Meizu'
酷派	'Coolpad'
乐视	'LeEco'
联想	'Lenovo'
'HTC'	'HTC'

Database Connection:

To download the gender_age_train and phone_brand_device_model onto python we have connected through SQL by using MySQL package "mysql.connector" in python.

host-	'cpanel.insaid.co'
user-	'student'
password-	'student'
database-	'Capstone1'

To download the events_data onto python we have to download the data set by using pandas package pd.read_csv("events_data.csv").

Explanation of tables and columns

gender_age_train:

This size of the data set is 74645*4. That means we have 74645 entries and 4 columns.

The columns are device_id, gender, age and group.

device_id is the unique id. The data type is int.

Gender is an object type having 2 values M and F.

Age is of int type ranging from 1 to 96 having an average age of 31 years.

The group is object type having ranges of age for male and female.

phone_brand_device_model:

it has 87726 rows and 3 columns.

This data has the following column: device_id, phone_brand,

device_model.

Device_id is of int data type and data type of phone brand and device model is an object type.

This data tells us different brands of phone and their model for particular device_id.

We have multiple phone brands but for ease of understanding considering only the top 10 brands which are: Samsung, cool, oppo, vivo, huawei, xiaomi, meizu, htc, lenevo, LeEco.

events_data:

- It has 3252950 rows and 7 columns.
- The columns are event_id, device_id, timestamp, latitude, longitude, city, state.
- The data gives us information about the frequency of events on a particular device id based on locations like city and state. Timestamp column gives us the time of occurrence of an event. Latitude and longitude give us the exact location.
- Event_id, device_id, longitude, latitude have int and float data types while timestamp, city and state have an object data type.

Summary of data mining

Events_data:

- 377 Null values in states which is about 0.01% of total data, this may change if filtered on Maharashtra state only.
- 423 Null Values in Lat & Lon data which is about 0.013% of total data.
- 453 Null values in device_id data which is about 0.014% of total data.
- After filtering the data of Maharashtra state:
- Rows reduced to 677105.
- Missing values reduced from 453 to 72 in device_id.
- Missing values in Lat & Lon reduced from 423 to 63.
- Missing values in the state reduced from 377 to 0.
- We have outliers in longitude and latitude which is wrong information when plotting.

Device Users Gender Age:

o we have outliers in age column which are 1 and 96.

phone_brand_device_model:

o most of the brands and model names are in the Chinese language.

Challenges in merging tables

Resolution:

o Most of the null values removed when we filtered out only Maharashtra data.
o Rest missing values is removed by using a simple imputation method.
o Gender_age_train: Outliers are removed using isolation forest.

phone_brand_device_model:

o We made 2 different lists of a brand in Chinese and brand in English. Then we zip those lists and Chinese words in phone_brand data are converted to English using lambda function.
o To resolve the challenges in merging table we did inner join taking device id as primary key on df_dev_model,df_mh and df_dev_model,df_events and we did left join on df_left_merge,df_gen_age_train,df_left_merge_all,df_gen_age _train taking device_id as primary key.

Proposed solution for the customer

o There are 77 distant phone brands. The maximum number of users use Xiaomi, closely followed by Samsung & Huawei.

These 3 phone brands account for ~60% users. The 7 phone brands (Xiaomi, Samsung, Huawei, vivo, OPPO, Meizu, Cool) account for ~90% users. These top 10 brands have 98.5% users in Maharashtra. All other phone brands contribute less than 1.5% each.

o We have 66.90% Male and 33.10% Female users. Age of maximum users (Male & Female) are around 20 to 40. Almost 50% of the users are of age between 22 to 32.

Solution:

For many people around the globe, especially women in developing nations, owning a mobile phone can mean the following:

-Financial independence.

-Employment.

-Better family health and education.

-Access to the Internet, government services, and information.

Insaid Telecom services can encourage women to migrate to the digital platform by announcing lucrative deals and offers exclusive to female customers.

o Mumbai and Pune have highest users as compared to other cities.

Solution:

Insaid Telecom Services can consider expanding to other districts of Maharashtra and diversifying portfolios in Mumbai and Pune.

o Male users are the dominant users across the top 10 phone brands. It is also observed that the top 10 phone brands are the same across gender. Both male & female users prefer Xiaomi as their top brand.

Solution:

Insaid Telecom Services can have exclusive SIM package deals.

- The maximum calls happen from 7 AM till 10 PM in the night.
- Hourly call distribution is almost similar between Pune & Mumbai (Top 2 cities) except between 6-7 AM Number of events between 6-7 AM is almost doubled than Mumbai. It looks like Pune users are early morning callers.

Solution:

Insaid Telecom Services can consider the creation of an app which measures the daily login time of a SIM holder. If a customer were to login at a particular time and maintain a streak, he/she can be awarded with an enhanced data usage limit.

Conclusion

As part of this project, we have studied the InsaidTelecom user and device usage data for Maharashtra. After doing the data analysis here are the following observations:

- In terms of mobile usage, Delhi is the topmost state. However, in terms of mobile users it is Maharashtra.
- We have observed that the age group between 20-40 are most users.
- We have more male users. Approx 66% as compared to 34% female users.
- In Maharashtra state, more than 90% of users belong to Mumbai and Pune. There are very few users in other cities, there are 87 cities in Maharashtra. InsaidTelecom can take other cities to increase their userbase.
- Xiaomi, Samsung, Huawei, Vivo are the most favoured brands of phones.