**Consulting Report**

**Table of Contents**

1. Introduction
2. Project Description
3. Problem Statement
4. Problem Analysis (Strategy for the Problem Statement analysis)
5. Sources of Data (Explain about database connection, tables, and their columns)
6. Summary of Data Mining (What kind of challenges you faced with the Data and how you resolved them? Summary of your Analysis)
7. Proposed Solution for Customers (Describe your Analysis in Detail)
8. Tools

* DS Tools
* Web UI Tools(You don’t have to explain this section)
  + PHP
  + JavaScript

1. Conclusion
2. **Introduction:**

**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.

1. **Project Description:**

We, at **Group 1003, are building** a dashboard to understand users' 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. Our Focused States are **Arunachal Pradesh, Chandigarh, Manipur, Tamil Nadu, Tripura, Uttar Pradesh.**

1. **Problem Statement:**
2. **Business Problem Statement**: User behavior across the focused State, which is going to directly impact the company's offerings. To deploy the right way forward and suggest actionable insights from marketing and product terms.
3. **Dataset Problem Statement:** We faced challenges with missing data in geolocations, device ids, handset names in Chinese language in the database, multiple tables with segmented data which were very important for getting the right insights. We were required to get a consolidated, curated data set and then apply the EDA steps to generate insights.
4. **Problem Analysis** (Strategy for the Problem Statement analysis)**:**
   1. **Understanding the Problem statement** - Understanding the activity that InsaidTelecom data project is given to us, which is the key to ensure the result is achieved with collaboration of the different Team members. Before thinking about the data, we have long discussions with the team about process, responsibility, timelines etc. This was the kickstart of our data initiative.
   2. **Getting data in python Notebook -** We have connected with MySQL to get the data for gender and Brands and upload CSV file for event data.
   3. **Exploring and Cleaning data -** We have started digging the data to see what we got and how we can link everything together to achieve our original goal. We have started taking notes on our first analyses and asked questions within the team to understand what all our variables mean. We started cleaning the data like filling the missing values, converting float into integer, converting absolute values, changing names from Chinese to English etc.
   4. **Enriching Dataset -** Now that we have clean data, it’s time to manipulate it to get the most value out of it. We have started the data enrichment phase of the project by joining all your different sources and group logs to narrow our data down to the essential data series. We have enriched our data by creating merged dataset, converting timestamp into date and time, extracting focused state for consulting etc.
   5. **Building helpful visualizations (EDA) -** Now we have a nice dataset, we started exploring it by building graphs. Visualization is the best way to explore and communicate our findings and is the next phase of your data analytics project. This helps us to enrich our dataset and develop more interesting findings. For example, by putting our data points on a map we could perhaps notice that specific geographic zones have more usage/users/brands than specific states or cities.
5. **Sources of Data** (Explanation about database connection, tables, and their columns) :
6. **Connecting with MYSQL database server to get the gender and phone brand files.**
   1. **gender\_age\_train** - Devices and their respective user gender, age, and age groups
   2. **phone\_brand\_device\_model** - device ids, phone brand, and phone models: note that few brands are in Chinese
7. **Downloading the CSV file for event data**
   1. **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 frequency of mobile usage. timestamp: when the user is using the mobile.
8. **Summary of Data Mining** : (Challenges faced with the Data and how we resolved them. (Summary of Analysis)

|  |  |  |
| --- | --- | --- |
| **SN** | **Challenges** | **Resolution** |
| **1** | Missing values in State data series | Replaced with mode with reference to City data series |
| **2** | Missing latitude / longitude data series | Replaced with mode with reference to City data series |
| **3** | Negative values for Device id data series | Made all the device id positive using absolution function |
| **4** | Data type of device id is found to be float in CSV data, whereas it is int in SQL files | Converted device ids in all the database as int64 |
| **5** | Timestamp contains both date and time | We have split the date and time with 2 different data series |
| **6** | Chinese names in Phone Brand data series | Converted into English names with reference to respective Chinese name by creating dictionary |
| **7** | Merging data set | We have merged data taking device id as reference and using merge function with “inner”. |

1. **Proposed Solution for Customers** (Description Analysis in Details) :

We have analyzed the data based on following EDA points:

|  |  |  |
| --- | --- | --- |
| **SN** | **Exploratory Data Analysis** | **Focused State** |
| 1 | Distribution of Users(device\_id) across States. | **Highest** no. of Users (device\_id) found in **TAMILNADU** with **44 nos.** |
| 2 | Distribution of Users across Phone Brands (Consider only 10 Most used Phone Brands). | **Highest** no. of Users is having **SAMSUNG** as Phone Brand with **22 nos.** |
| 3 | Distribution of Users across Gender. | **43 nos. of MALE** usersand **27 nos. of FEMALE** usersfound in our consulting States. |
| 4 | Distribution of Users across Age Segments. | **Highest** no. of users is coming under the age segment of **M 39+** with **10** unique users. |
| 5 | Distribution of Phone Brands (Consider only 10 Most used Phone Brands) for each Age Segment, State, Gender. | **Age Segment: SAMSUNG** is the most usedBrand **(4 nos.)** which is present under the Age segment **FEMALE 27-28.**  **State: SAMSUNG** is most used in the **State TAMILNADU** with **13 users.**  **Gender: SAMSUNG** is used by most of the **FEMALES** gender with **13 users.** |
| 6 | Distribution of Gender for each State, Age Segment and Phone Brand (Consider only 10 Most used Phone Brands). | **State: MALE in TAMILNADU** has the **highest gender** users with **26 nos.**  **Age Segment: M39+** is the most used **Age Segment** with **10 users.**  **Phone Brand: SAMSUNG** is used by most **FEMALE with 12 users** |
| 7 | Distribution of Age Segments for each State, Gender, and Phone Brand (Consider only 10 Most used Phone Brands). | **State: TAMIL NADU** has the highest no. of users across all the focused States in the age segment of **39+** with **7 users.**  **Gender**: **MALE** in the age segment of **M39 +** is having the highest no. of users with **10 users**  **Phone Brand:** **SAMSUNG** has the highest number of users in focused States with **4 users** for the age segment **F27-38.** |
| 8 | Hourly distribution of Phone Calls. | **Highest** no. of users in focused States is found to be active between **11 -12 pm** with **290 events** |
| 9 | Plot the Users on the Map using any suitable package. | We have found few latitude and longitude data outside the region in **Tamil Nadu** State |
|  |  |  |

1. **Tools :** DS Tools used

We have used the following DS Tools

1. Jupyter Notebook / Google Collab
2. Python (Library : Panda, Pandas\_Profiling, Numpy, Matplolib, Seaborn, Sweetviz,Folium, Geopandas)
3. SQL and Excel
4. **Conclusion :**
   1. **TAMIL NADU** has the most users in both **MALE and FEMALE.** We recommend clients to focus more on **TAMIL NADU** in our state group.
   2. **SAMSUNG** is the most used brand in focused States. Clients can make combo offers of their call plan with new users who purchase SAMSUNG brand mobile.
   3. **Most people** are active between 11 -12 pm in most of the States. Clients can run the promotions activity in this hour.
   4. **M39+** is the Age segment with maximum users. Clients can propose unique propositions for the segment. We can promote OTT content like Netflix, Hotstar, Amazon Prime etc. to generate additional revenue.