Data Analysis and Visulization CS306

Instructor : Prof. Pankaj Kumar

Introduction

- The dataset on which we had worked was "yes_bank.csv". It consisted current and permanent addresses of yes bank customers.

- The data has 9 columns which consisted of customer_id_token, 4 fields of current address, consisting of current address column, current city, current state and current pin code and similarly 4 fields of permanent address.

- The dataset consisted of 10594391 rows.

Data Cleaning Techniques

- The data consisted of many garbage entries in which there would be numeric entries(0,1) instead of address string.
- Then there would be strings like "dummy" and "add address" and "add city" which need to be removed.
- So we searched such strings in the dataset and removed such entries.

Handling Missing Data

- We observed that there were several entries in which the address line field was empty. This was because the data provided was of address line 3. In many cases customers do not enter address line 3, and write their complete address in only address line 1 and 2.

- So it would be unfair to remove such entries as that customer is actually valid, so we have kept that entry.

- We checked that if city, state, pin code of either current or permanent type is null, then we have removed that entry.

Outlier Detection

- We have used a python library geopy to find the latitude and longitude of cities. We were unable to find the latitude and longitude of all the cities. So the places of which we were unable to find, were classified as outliers.

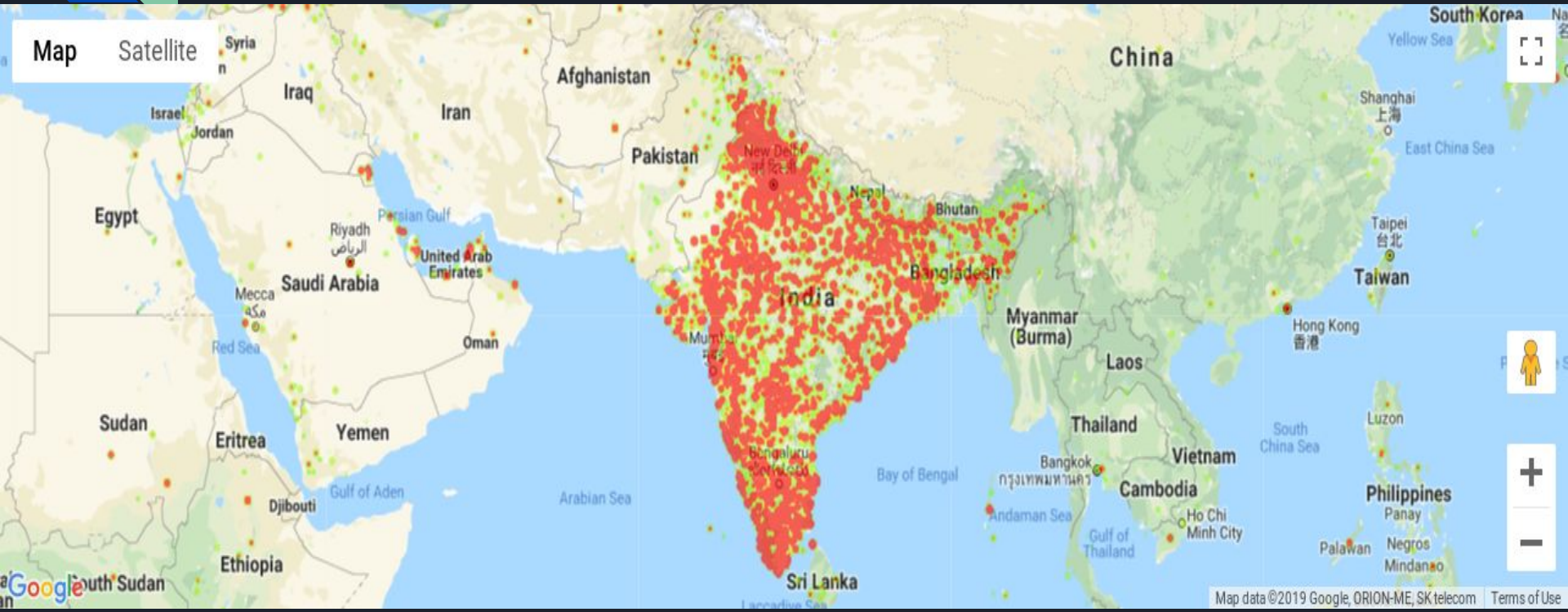
- We were able to locate the addresses of nearly 59 lakh customers and showed it on the heatmap. So the remaining 41 lakh are outliers.
- Data normalization and data augmentation were not possible on this data set.

Procedure for Heat Map generation

- One of the methods was to randomly select some data(say 1 million rows) and work on that data. But the method described below allowed us to work on more data.
- What we did was we visualized the customers based on their permanent cities.
- The data available in the address field was not complete and so a lot of times was not recognized by the library. However this happens fewer times with city.
- Number of unique cities were nearly 1 lakh. We found the exact latitude and longitude of this cities using geocode() function of geopy library. We found locations of 11,000 permanent cities and then correspondingly found in the cleaned dataset the entries which have these permanent cities.
- We found nearly 58 lakh entries. Then we found the frequency of each city and then made the heat map of permanent cities.
- For current city we did the same thing, only change being we found 59 lakh entries.

Heat Map of world based on permanent city

Heat Map of India based on permanent city



Heat Map of world based on current city

Classification of customers

- The aim was to develop a technique in order to classify the given customers into business or household based on the data given.

- Since a large size of the data given was incomplete, it was not possible to use any machine learning technique.

- We did two things. Firstly we compared the current and permanent addresses given in the dataset and if they are not equal than that customer is classified as business one.

Classification of customers

- From the remaining customers whose addresses are equal, there is possibility of being a business one. We analyzed the data and saw if the address field consisted of certain words such as "MALL", "OFFICE", "COMPLEX", "CHAMBER", "BAZAR", "CENTRE" ,"BANK", "SCHOOL", "DAIRY", "POST" ,"FACTORY","ESTATE", "MARKET","HOTEL","SHOP""FLOOR","COMMERCIAL","HOSPITAL","COLLEGE","UNIVERSITY","METRO","SALES","SCHOOL","HALL","INDUSTRY","INDUSTRIAL" and didn't consist of certain words such as "NEAR","BEHIND", "NR", "OPP"; than that customer would be classified as business.

Classification of customers

- The reason behind doing this was that the first set of words belong to some sort of business buildings and not a household one; and the second set of words if present, would mean that the address is somewhere around a business complex, and so is a household one; so second set of words should not be present.

- The remaining set of customers are classified as household. We were able to classify 2 lakh customers as business and 18 lakh customers as household. Heatmaps of both are shown below.

Heat Map of world based on business addresses

Heat Map of world based on household addresses

Thank you