Exploratory Data Analysis of Mobile Data Speed To Determine The Best Service Provider In India During 2018-2019

Ayush Porwal
Dept. *Chemical Engineering*

Birla Institute of Technology and Science, Pilani - Hyderabad Campus
Hyderabad, India
f20171090@hyderabad.bits-pilani.ac.in

Azizur Rehman
Dept. *Mechanical Engineering*

Birla Institute of Technology and Science, Pilani - Hyderabad Campus
Hyderabad, India
f20170674@hyderabad.bits-pilani.ac.in

Mayank Negi
Dept. *Computer Science Engineering*

Birla Institute of Technology and Science, Pilani - Hyderabad Campus
Hyderabad, India
f20180210@hyderabad.bits-pilani.ac.in

Hari Kiran
Dept. Civil *Engineering*
Birla Institute of Technology and Science, Pilani - Hyderabad Campus
Hyderabad, India
f20171462@hyderabad.bits-pilani.ac.in

***Abstract*— It is considered that the speed of the internet connection is the most essential and fundamental metric to measure the user's internet experience. In our study, we choose a crowdsourced dataset from 'https://data.gov.in/'. The data collected may be slightly biased since it is very likely that the user will initiate manual tests when he experiences the poor performance of the mobile network. The aim of this project is to extract useful information by applying suitable techniques of data mining like data cleaning, data preprocessing, data visualization on the given dataset and extract reasonable correlation of provided attributes and mold it into an understandable structure for future usage.**

Keywords—data mining, data preprocessing, mobile speed data, and data visualization.

I. INTRODUCTION

The telecommunications industry has been one of the first to adopt data mining technology. This is because telecom industry generates and stores enormous amounts of data containing important parameters of operation in the network. However, the Industry also faces a number of data mining challenges due to the enormous size of their data sets, the sequential and temporal aspects of the data. Hence, the need to predict important correlations is essential for the overall improvement of user satisfaction and service performance.

Exploratory data analysis was performed on a processed dataset to analyze the different relationship between the attributes present in the data set. Data preprocessing techniques such as sampling, normalization, feature creation, and data cleaning were applied to the dataset. Suitable visualizations were then made between those attributes as well. The raw data set is a crowdsourced collection of measurements from all over India.

II. DESCRIPTION OF THE DATASET

A. In this paper, the dataset being worked upon is a crowdsourced measurements of mobile data speed provided by the National Data Sharing and Accessibility Policy (NDSAP) on their website (*data.gov.in).*

B. Data Contributors: Ministry of Communications, Department of Telecommunications (DOT), and Telecom Regulatory authority of India (TRAI).

C. Dataset Information:
   - **(i) Type:** Multivariate.
   - **(ii) Number of instances:** 12649245
   - **(iii) Attribute Characteristics:** Categorical and Numerical (int)
   - **(iv) Number of attributes:** 7
   - **(v) Objects with missing values:** Yes

D. Attribute Information:
   - **(i) Operator:** Jio, Airtel, Idea, Vodafone, Cellone, Uninor, Aircel, Dolphin.

(ii) **Technology:** 2G, 3G, 4G.
(iii) **Test Type:** upload / download.
(iv) **Data Speed:** initially measured in kilobytes per second(kbps) but normalized during data preprocessing.
(v) **Signal Strength:** Decibels with reference to one milliwatt (dBm).
(vi) **LSA:** Licensed Service Area.

III. DATA PRE-PROCESSING

This Technique was used to convert the raw data (consisting of missing values, unorganized etc.) into useful form of data, which as a result helps in proper data analysis.

The sequence followed was:

(a) Concatenation of data from multiple .csv files to a single excel file by the addition of a new feature 'Month&Year'.

(b) Random sampling without replacement was done as the dataset was too large. The sample was 70 % of the size of population. It is still very large.

(c) The addition of new column resulted in the creation of a new feature(Month&Year).

(d) Then, the merged data was cleaned by both removing the missing value rows for the feature Signal_Strength and forward filling the data for the column 'LSA'.

(e) The column 'Data_Speed' was normalized (Data Transformation) as the range was too wide. As a result, the normalized data speed would vary from 0 to 1.

(f) Since, the number of features were less, there was no requirement for dimensionality reduction.

Statistical Analysis:

1. **Outlier Detection**:
   a. Number of observations classified as outliers: 458560.
   b. Number of observations classified as non-outlier: 12190685
2. **Covariance** was found to be: 0.701932784104611
3. **Pearson Correlation** was found to be: 0.22772344947987816
4. **Spearman Correlation** was found to be: 0.35040558990649073

IV. VISUALIZATION OF THE PROCESSED DATA

After the anticipated preprocessing was done, Initial visualizations were made in the form of box plots, bar graphs, pie charts, scatter plots to better understand the processed data. The plots are described below with their following observations.

(i) Scatter plots for each operator were made between signal strength and normalized data speed [0-1] across all broadcasted spectrums. There was no linear relationship between the two.

(ii) Bar graph of Normalized average data speed across all broadcast spectrums for each operator was made. The order of Average data speeds was 'Jio>Airtel>Idea>Vodafone>Uninor>Cell one>Dolphin>Aircel'.

(iii) Bar graph of Normalized average data speed across all broadcast spectrums for each state (LSA) was made. The highest Average data speeds were seen in Bihar and the lowest being the North east.

(iv) Pie chart consisting of the technology distribution was made across all states. The values were 4G:91.41%; 3G:8.58%; 2G:0.01%.

(v) Pie chart consisting of test type distribution was made across all states and spectrums. The values were Upload:50.06%; Download:49.94%.

(vi) Pie chart consisting of operator distribution across all spectrums was made. The values were Jio:63.10% Airtel:17.47% Idea: 9.39%, Vodafone:9.85%, Cellone: 1.68%.

(vii) Scatter plot consisting of signal strength range across all spectrums for each LSA was made. The range was [-140,0] dBm. The higher is the value, the better is the signal strength.

(viii) Box plot consisting of Data speeds for upload as well as the download was made across all spectrums for each LSA.

(ix) Box plot consisting of Data speeds was made across each spectrums for every LSA.

(x) Area chart consisting of Variation in Average data speeds for each spectrum over the time frame of march'18 till july'19.

(xi) Geo spatial chart consisting of average signal strengths for each LSA over the captured time frame which is march'18 till july'19.

(xii) Normalized minimum and maximum data speeds of each operator for every LSA was made over the observed time frame.

(xiii) Geo spatial chart consisting of average data speeds for each LSA over the captured time frame which is march'18 till july'19.

(xiv) Trend of average data speed with respect to state, technology and time.