ISP Recommender and Analyser System

P Yedhu Tilak
2017A7PS0021H
f20170021@hyderabad.bit
s-pilani.ac.in

Prakhar Gupta
2017A7PS0121H
f20170121@hyderabad.bit
s-pilani.ac.in

Akhil Agrawal
2017A7PS0190H
f20170190@hyderabad.bit
s-pilani.ac.in

Smit D Sheth
2017A7PS1666H
f20171666@hyderabad.bit
s-pilani.ac.in

Abstract

With the rapid growth of the telecommunication sector in India and an increased number of telecom providers, it becomes important for an individual to choose the best service provider suitable in the area in which he/she resides. Due to heavy competition between the service providers, it also becomes important for them to know the regions in which they need to improve.

A data mining model with regression and clustering will solve both of the above problems. Regression techniques will help us predict the data speed values for a few given features such as service provider, technology, signal strength, and service area. Clustering will allow us to know the regions in which a particular service provider is better. This paper presents the details of data cleaning (filling missing values and outlier removal), data visualization and data pre-processing (standardization, stratified sampling, and one hot encoding). Applying these methods will help in the regression and clustering techniques to be used further.

Keywords—data mining, machine learning, regression, data speeds, signal strength, outliers, standardization, one-hot encoding, data visualization

I. Introduction

In the past decade, there has been significant growth in the Indian Telecommunication sector. With India progressing towards a perfectly competitive market, consumers face a serious challenge of choosing the most appropriate mobile operator for them. Even service providers need to know in which regions they need to improve their service and reachability.

In a country like India which was recently told to be ranked 1 in the world in data usage per smartphone per month, Data is abundant and Data Mining is essential.

The first problem we have identified is the task of predicting the data speeds of users based on their signal strength of a particular service provider in the given region. This will help the consumers estimate the data speed for them.

Secondly, we are going to provide a solution to the users in the form of a recommender system that helps them choose the most appropriate service provider in a particular area on the basis of criteria such as signal strength, speed, etc.

Thirdly, we are also providing the solution to the service providers by predicting which regions need strength boosting based on the overall distribution of signal strength across India.

To tackle such problems, we are going to apply various data mining techniques such as regression, clustering, etc and based on inferences from the visualizations obtained we are going to providing the best solutions to the users.

II. Background

The number of people connected to the internet is growing day by day. This has led to the rapid growth of the cellular service market and various companies have entered it and are competing against each other. This competition has brought upon much needed choice in the sector and forces the cellular companies to be at their best. Due to this choice, various users have a task at their hand of choosing which cellular service is best for them, and companies should work on staying competitive.

Till recent times, these tasks were solved using people's opinions and feedback. Earlier, A User would choose a service based on recommendations and other people's opinions, and companies would improve based on feedback. But with the advent of Machine learning and Data mining, both of these two tasks can be automated. Various models can be built that can suggest a user which cellular service he should go for, what can he expect. Models can also be built to tell each cellular service in which areas they are lacking, and in which areas they are flourishing.

One of the main challenges that will be faced is **data**. Without proper, adequate and accurate data, no model can produce good results. But in the modern era, where storage technology is at its best and various modern sensors and data collection techniques ensure that a large quantity of accurate and precise data is readily available.

Another major encouragement to take this approach is the development of effective and efficient Data Mining and Machine learning algorithms. Various algorithms can be employed which can extract inference, make predictions and correct errors much faster and efficiently than any human or normal software. These algorithms can be used to train our models to perform better.

All of these points point in the direction of solving these huge problems using Data mining and Machine learning.

III. Methodology

To tackle these problems, we have employed a full machine learning and data mining pipeline, right from data collection, cleaning, visualization to analysis and inferences. Various techniques are used in each stage to make the development and analysis process more effective and efficient, and reach final goal to build models that help us in solving these problems. Various software tools such as numpy, seaborn, etc were used to make our task simpler. (Seaborn, n.d.)

A. *Dataset*

We have worked on the **All India Crowdsourced Mobile Data Speed Measurement for July 2019** dataset provided by the Indian government on the website https://data.gov.in/.

URL for the dataset: https://data.gov.in/resources/all-india-crowdsourced-mobile-data-speed-measurement-july-2019
This dataset provides mobile data speed and signal strength for all the states as well as few cities of India for the month of July 2019. The information provided in the dataset is measured using the TRAI MySpeed app.

The dataset consists of more than 6,00,000 user mobile service data points for the month of July 2019. In detail, there are 6 feature columns: **service provider, technology (3G/4G), download/upload type, data speed (in Kbps), signal strength and service area**. The data speed and signal strength are continuous attributes, while the rest of the features are categorical. We have observed a satisfying correlation between signal strength values and data speed values. This will help us to predict data speed values for other known attributes.

We have observed that around 11.17% of the signal strength values are missing and around 1.74% of the service area data is missing. We have come up with a solution to fill these missing values. It is briefly described in 3.2.1 Filling missing values section of this report. To reduce the size of the dataset, we have used stratified sampling.

B. Data Cleaning

Data cleaning is also one other important aspect of any machine learning task. This involves performing various operations that make our data more consistent and better, prepare them to be analysed. This step is very important as Data that is not clean/inconsistent might lead to different, and sometimes negative results. The following tasks were performed to clean the data:

Filling missing values

The dataset for the month July 2019 has a large amount of missing values for the signal strength attribute (about **11.17%** of all signal strength values) and a few missing values for the service area field (around **1.74%**). Rest all the feature attributes have no missing values at all.

The rows having missing values for the service area attribute were simply dropped because they were present in a very small fraction. Due to their presence in a small ratio, they made negligible contribution to the overall distribution of the data points. Therefore, dropping them was a good idea.

For filling the missing values in the signal strength column, the following steps were followed:

1. Groups having the same service provider, technology, download/upload type and service area were formed. The reason to create groups and not to work on the dataset as a single entity was that the distribution of the signal strength varied among these 4 attributes. Considering the dataset as a single entity would violate this underlying distribution.

2. We calculated the median of each group with the help of the data points belonging to that particular group. Due to the presence of outliers in the dataset, we used median and not mean. The mean would possibly give erroneous values due to outliers.

3. The missing values were filled by the median of the group in which the rest 4 features of that row belonged.

The above steps were performed with the help of a 4-dimensional dictionary in Python. Each dimension represented one out of 4 features mentioned above.

After performing the above steps, still there were 31 missing values in the signal strength column. This was because very few of the groups created had missing values for all the data points present in that group. All the 31 rows having missing values (around 0.005%) were dropped.

Finally, there were no missing values in the dataset.

Removal of outliers

An outlier may be due to variability in the measurement or due to experimental error. This can cause serious problems in our statistical analyses and regression models.

We have used Box plot as a visualization tool to study outliers. It is a method for graphically depicting groups of numerical data through their quartiles. The individual points above the upper whisker or below the lower whisker can be considered as outliers.

We have used our regression target variable Data_Speed_Kbps to make the box plot. Outlier Removal is done separately for upload and download data points because of significant difference in their mean data speed. Data points belonging to some of the states are not considered in outlier removal as their box plots indicated no outliers.

A large percentage of points were found above the upper whisker and all of those cannot be outliers. Thus, we calculated the threshold using the formula **q75 + k*(IQR)** where k is optimized to remove 1% of data points.

q75 - 75th percentile
q25 - 25th percentile
IQR -Inter Quartile Range- q75-q25

C. Data Visualization

Data visualization is a key part of any Data mining or machine learning tasks. They help us to visually represent the underlying patterns and information of the data in a clean and simplistic way, from which various insights can be gathered quickly and accurately. The mobile speed data used in this project is mainly categorical and has only two continuous attributes (speed and strength) and the best way to represent categorical data is using bar plots, box plots and histograms.

Box plots are extremely useful in visualizing distribution of data over certain attributes and in detection outliers, as explained in the cleaning section. To get a brief idea about which values are important and occur more frequently, we plotted bar graphs with values of an attribute on one axis and the count of the number of occurrences in data on the other axis. We used this technique of mainly 3 attributes, name Service_provider, technology and location. These plots give us a brief idea on which service provider is most popular, which cellular technology (3G or 4G) is used most, from which location the most number of samples were gathered, etc. using these graphs we were able to say that JIO is the most popular service, most people use 4G, etc.

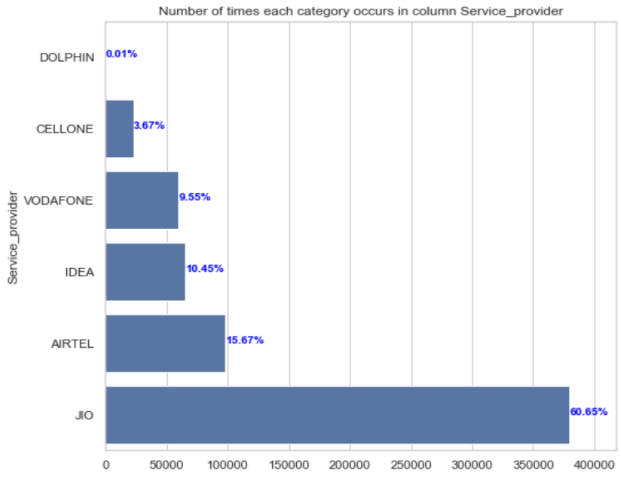

Fig 1: Bar plot of count vs Service_provider

Bar plots of mean data speeds vs location and mean data speeds vs Service_provider was drawn, which gave us various insights about which service provider provides the highest and lowest speeds, which locations are benefitting from high average speeds, etc. This also gives us an initial idea on which service provider is lacking in terms of service and which service provider is performing well in general.

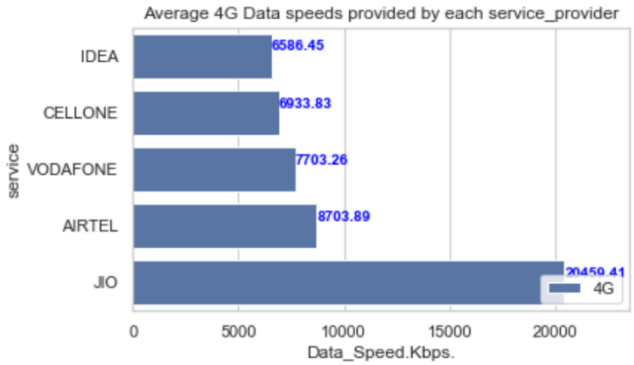

Fig 2: Average speed for each service provider

Finally, we have plotted a 2-D histogram of Data_Speeds and Signal_strength. This plot shows us how Data_Speeds vary with respect to signal strength and relative frequency of points that lie in each region of the graph. This also shows that there exists a healthy correlation between speed and signal strength.

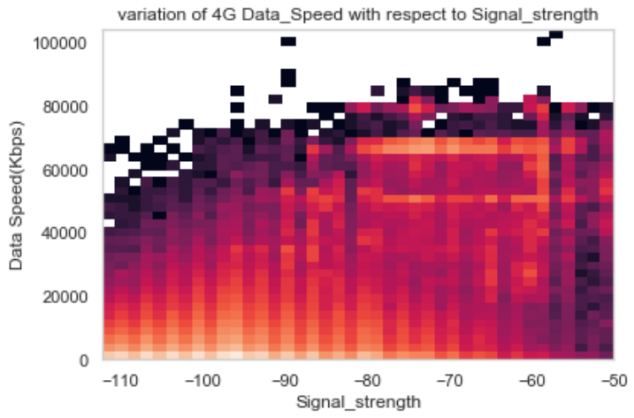

Fig 3: Data_Speed vs Signal_strength

D. Data Preprocessing

Data pre-processing involves modifying our data to bring it to a form more suitable to our needs. Certain pre-processing techniques improve our results and certain others speed up the analysis process. Pre-processing can include data transformation, reduction, etc. Various pre-processing techniques were applied that suit our analysis methods the most.

Stratified Sampling

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. It increases the speed and accuracy of study and significantly reduces the cost.

In our dataset the number of data points corresponding to each (state, provider) tuple varies significantly. Thus, we are going with Stratified Sampling. Data points corresponding to each (state, provider) tuple constitute a stratum. Sampling from each stratum is performed in proportion to the strata size.

Inside a stratum, points are collected using simple random sampling

The formula used to calculate the number of sample points in a stratum is:

$(N_i / N) \times (n)$

N_i= number of points in the strata
N= total number of data points
n= Total sample size

Standardization

The only two feature columns with continuous values, namely data speed and signal strength, were standardized. Standardization improves the quality of the data. The formula for standardizing every entry X of a column with mean μ and standard deviation σ is:

$$z = \frac{X - \mu}{\sigma}$$

After standardization, the distribution of the data speed values and the signal strength values had mean = 0 and variance = 1. This helps in two ways. One is that the data is scaled down from large values to similar and small values. This helps to reduce the complexity of all the algorithms to be used further on the dataset. Another advantage is that, the standardized distribution of the data points will help us find various relations between data speed or signal strength and other features. This will make the prediction of data speed values easy.

One-hot encoding

Most of the datasets come with nominal attributes which are categorical in nature. These categorical attributes have values which are basically text labels/strings. In this raw form, the computer/algorithm cannot extract meaning from them. So, the best approach is to convert them to some form of numerical attributes.

Since all of the categorical attributes we have are nominal in nature, the best to handle these is to make a new

feature, a binary asymmetric feature for each value of the categorical attribute. This process is called one hot encoding of categorical features. This process was performed on technology, Service_area and service_provider attributes of the dataset. This technique also increases the dimensionality of our dataset, which does not create a problem in our case as we ample amount of data points.

*E. **Data Analysis***

We are going to predict download and upload speeds by applying regression based on attributes such as signal strength, service provider, region, etc. Also, we are going to do cluster analysis-based attributes such as region, signal strength, etc in recommending the best possible service provider in a particular state.

These clustering and regression techniques we are planning to use work very well when the data provided to it are in specific form. So various cleaning and pre-processing techniques were applied to bring our data to this form. Missing values are faults in our data and they need to be corrected or else we might lose some value data. Outliers tend to pull the regression line towards them, skewing our results, hence these were removed so that we can get the best possible line. Stratified sampling was performed on our data as we had a very high number of data points, and reducing them properly speeds up our process. Standardization was performed to bring all the continuous attributes to the same scale. Regression also assumes that the underlying data follows a normal distribution and hence transforming our data to this form helps. Finally, one hot encoding converts categorical attributes into computer understandable form.

IV. Results and Discussions

Before the missing values were filled / removed, we had around 11.17% missing values in the signal strength column, and 1.74% missing values in the service area column. Our dataset now has 0 missing values in all the columns, while the distribution of the data points remains unaffected.

Outlier analysis resulted in removal of around 5856 data points. The upper threshold for download data points was set to 74951 kbps and around 2169 points were identified as outliers. Similarly, the upper threshold for upload data points was set to 19940 kbps to identify and remove 3687 outliers. The threshold values were set to remove around 1% of the total data points as outliers.

Various data visualization techniques were applied to get an initial idea of the underlying patterns of our data, and some interesting findings are obtained.

Most of the users are on 4G networks, around 95% and the rest 5% are on 3G networks. Jio is the most popular cellular service, having a market share of 60.6%, followed by Airtel. Dolphin is the least popular network, used by only 0.01% of the users.

Maharashtra has the highest number of users, whereas Jammu and Kashmir and the North-east states have the least. Highest average data speeds are obtained in Delhi, which can be due to fact it is the capital of our country. Jio provides the highest average speeds, which explains why it is popular among the users.

The dataset size reduced to around 6.4 lakh points after data cleaning and outlier removal, in order to further increase the accuracy and speed of our analysis we performed stratified random sampling, points corresponding to each (state, provider) tuple formed a stratum. Keeping the total sample size as 50% of the dataset size, we sampled around 3,244,78 points.

Standardization performed on the continuous value attributes, i.e., data speed and signal strength made the distribution of data points of both the columns standardized with mean = 0 and variance = 1. This also scaled the values of both columns to a small and uniform scale.

One-hot encoding as explained earlier, is used to convert categorical attributes to asymmetric binary attributes. This technique increased the dimensionality of our dataset from 6 to 35.

V. Conclusion

The primary aim of this project is to provide an approach where we can make the lives of users easy, by providing automated techniques using which they can select the best cellular service for them, and what they can expect from each cellular service. Companies can also benefit from our models as they help them in identifying the areas in which they should improve.

All of these are possible due to the recent growth of machine learning and the ability to gather large amounts of accurate data. We have employed various data cleaning, pre-processing and visualization techniques to prepare the data and bring it to the right form. On this data, we will employ various regression and clustering techniques to develop models that will help us to solve these problems.

VI. Bibliography

Seaborn. (n.d.). Retrieved from Seaborn: https://seaborn.pydata.org/

Sharma, N. (n.d.). *Detect and remove outliers*. Retrieved from Towards Data science: https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba

Trek, S. (n.d.). *Stratified Sampling*. Retrieved from Stat trek: https://stattrek.com/statistics/dictionary.aspx?definition=stratified_sampling