Call Quality Analysis

Drishti Mamtani
Department of Computer Science
BITS Pilani Hyderabad Campus
(BITS Pilani)
Hyderabad, India
f20160574@hyderabad.bits-pilani.ac.in

Himanshu Sharma
Department of Computer Science
BITS Pilani Hyderabad Campus
(BITS Pilani)
Hyderabad, India
f20160886@hyderabad.bits-pilani.ac.in

Shreyansh Jain
Department of Electrical and Electronics
BITS Pilani Hyderabad Campus
(BITS Pilani)
Hyderabad, India
f20160947@hyderabad.bits-pilani.ac.in

Amartya Vats
Department of Electrical and Electronics
BITS Pilani Hyderabad Campus
(BITS Pilani)
Hyderabad, India
f20160628@hyderabad.bits-pilani.ac.in

***Abstract*— The paper deals with the analysis of the Call Quality dataset using operator, location, data speed, call drops and signal strength as the key attributes. The data pre-processing techniques used are removal of null values, min-max normalization, equal interval discretization, and outlier analysis. The techniques proposed for analyzing the dataset are multiple regression, logistic regression, clustering, and chi-squared test. These will help to draw various results like the best operator in a particular city, relationships between location, data speed, signal strength, and other significant features. Various pairs of features are also tested for dependence. These techniques give a holistic view of the features and their effects on the call quality.**

Keywords: Data speed, signal strength, chi-square test, min-max normalization, F Test, t-test, adjusted R square, silhouette value, root mean square error, null hypothesis, alternate hypothesis, regression, level of significance, observed frequency, expected frequency, tolerance level, logistic regression, clustering, internet service providers(ISPs).

I. INTRODUCTION

Call quality analysis is crucial for both internet service providers and customers. ISPs could analyse the market and understand their competitors' merits and flaws to enhance their own customer base. Whereas for customers, this would help them to choose the best option available. So it is a win-win situation for both the parties. The analysis can be divided into three major parts, pre-processing, understanding the dataset through visualizations, and data analysis algorithms. Each aspect is equally significant and none of them can work alone. Pre Processing includes transforming raw data into a useful and efficient format. Data visualizations include various plots and graphs to understand the dataset better. The algorithm techniques enable in applying various models on the dataset and get concrete results. In the paper, multiple regression is used to find the relationship between upload and download data speed, signal strength, and location. This analysis is done for the complete dataset as a whole, followed by analysis for each operator. Similar models are used to find relationships between data speed and signal strength.. K means clustering is used to predict the best network operator for a particular location. This would directly help the customers to choose the best network in their city. For a particular city, the latitude and longitude are found, which are used as parameters for clustering. Logistic Regression is used to find the relationship between the call drop category and other features. Chi-square tests will be used to find if literacy rate and network technology, state rainfall and rating, population, and rating are independent of each other or not.

II. DATASET

A. Dataset description

The dataset consists of details of numerous callers located in different parts of India. For each caller, different attributes like the network service provider, the network technology (4G or 3G) that the caller was using, in-out mode he/she was in that is whether he/she was traveling or not, what rating he/she gave to the call, etc. were collected. The dataset is obtained by combining a few datasets of September 2018 from data.gov.in. The dataset is a good mix of qualitative and quantitative features. The features are as follows: service provider(R-Jio, Vodafone, BSNL, Airtel, Idea), in-out traveling communication(Indoor, Outdoor, Travelling), network technology (3G, 4G), rating (1-5), Call drop status(satisfactory, poor voice quality, call dropped), latitude, longitude, state(Tamil Nadu, Punjab, Haryana, etc.), average upload and download data speed, average upload and download signal strength, population, rainfall, literacy rate and area of the state.

III. DATASET PRE-PROCESSING

Real-world data is often incomplete, inconsistent, and is likely to contain many errors. Data pre-processing is done to resolve such issues. Data pre-processing involves data cleaning, data transformation, and data reduction techniques. In this paper the focus is on only data cleaning and data transformation. Data reduction just involved combining two features namely signal strength upload and signal strength download to form a new feature called average signal strength. This would be covered in the next section of the paper.

A. Data Cleaning

There were many null values in the dataset. For some objects, the state was not mentioned and for some network was unknown. As the analysis is primarily based on the network providers and the state, the objects with many null values were removed to obtain consistent data. There were many objects with an attribute value of state as central region, Chandigarh, etc. So these tuples were removed and data of 15 states was left. After data cleaning, only two objects were left with 2G as the network type. Two objects alone cannot lead to appropriate results and hence they were also deleted. This also implied that India had shifted to advanced network technology namely 3G and 4G.

B. Data Transformation

After this, min-max normalization was performed to transform the data from the measuring unit to a new interval ranging from new_min to new_max.

$$new_{val} = \frac{old_{val} - min}{\max - min}(new_{\max -} new_{min}) + new\_min$$

Here min and max are old minimum and maximum values of the attributes. All the continuous features (longitude, latitude, average data speed download, average data speed upload, average signal speed download, average signal speed upload, population, area of the state, literacy rate and rainfall) were converted to a [0-1] range that is new maximum was 1 and the new minimum was 0. This was done so that features with larger values do not overwhelm the clustering and regression models. Next discretization with equal intervals was used. Discretization is the process of transferring continuous functions, models, variables, and equations into discrete counterparts. This was done for literacy rate, rainfall, and population so that these features could be used for chi-square tests at a later stage. After this outlier removal was performed. Outliers are the data points that are significantly different from the remaining dataset. An outlier can be a valid point or noise. Outlier analysis is done to ensure that the extreme cases do not impact the analysis thus, maintaining the accuracy of the model. Outliers were removed by calculating the z-score for all entries in continuous parameters (signal strengths, download speeds). Any entry having a z-score greater than 3 was considered an outlier and subsequently removed.

IV. Visualization

Data visualization is a graphical representation of information and data. Visual elements like box plots, correlation matrix, line histogram, bar plots, scatter plots, etc. are used to provide an accessible way to comprehend trends, outliers, and patterns in data.

Fig 1. Correlation Matrix

A. Correlation Matrix

Correlation analysis is used to quantify the degree to which the two variables are related. Correlation coefficients so evaluated depict how one variable changes with respect to other.

The following inferences were observed - There is a high positive correlation between average signal strength upload and average signal strength download implying that both should not be used simultaneously in a model. Strong negative correlation between the area of state and longitude indicates that as longitude increases the area of the state decreases. Also as one goes towards the east the amount of rainfall also increases, as the area of the state increases the population also increases and as the latitude increases the amount of rainfall decreases(since the observer moves away from the equator).

B. Box Plots

Boxplot is a way to display the distribution of data based on minimum, 1st quartile, median, 3rd quartile, and maximum values. It tells about the outliers and if data is skewed or is symmetrical. The following inferences were observed from the box plots - R-Jio has much greater average data speed

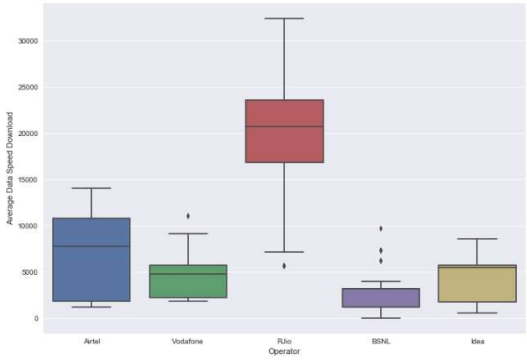

Fig 2. Box Plot representing Data Speed Download for different operators

download than other networks, signal strength is best for BSNL, data speed download and upload comes out to be best when outdoors and 4G Network is far more superior as compared to 3G and other networks in terms of data speed upload.

C. Scatter Plots

A scatter plot is a mathematical diagram used to plot data points on the XY plane in the attempt to show how one variable is affected by another.

The following inferences were taken -

1. Signal strength download, signal strength upload exhibit a linear relationship, and are highly correlated.
2. Weak relationship observed between average signal strength upload and average data Speed upload
3. A weak relationship was observed between average signal strength download and average data speed upload.
4. Weak relationship between average signal strength download and average data speed download observed.

D. Line Histogram

Line Histogram is used to observe the distributions of different parameters. These were drawn for average data speed download, average data speed upload, average signal strength download, and average signal strength upload. It was observed that since, average signal strength download and average signal strength upload had very similar distributions there was a possibility of a high correlation between them, which was further verified by the correlation matrix.

Some other interesting results were: It was observed that the majority of users/callers gave extreme ratings to Airtel. R-Jio, Idea, and BSNL had distributed ratings whereas Vodafone got good ratings. This showed that Vodafone won as a public choice. Majority of Airtel's users used 4G, Idea's user base had comparable 3G and 4G users, Jio only had 4G users, majority of BSNL users used 3G and had very minimal 4G users and Vodafone had a fair share of both 3G and 4G users.

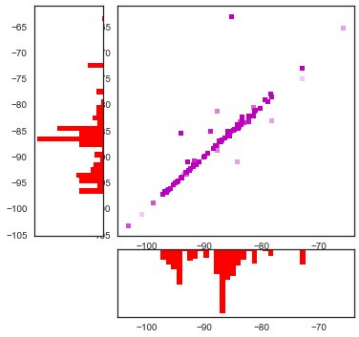

Fig 3: Scatter Plot representing a linear relationship between signal strength upload and signal strength download.

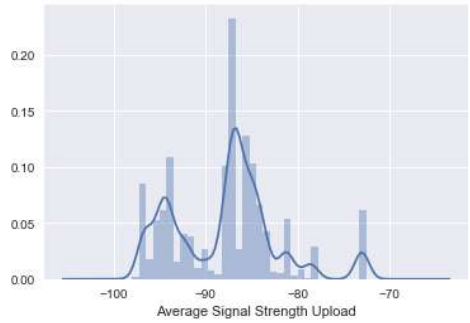

Fig 4: Line Histogram showing the distribution of

R-Jio is the most subscribed telecom operator among the users.

V. Data Reduction

From the above analysis it was clear that signal strength download and signal strength upload are extremely correlated hence, a third variable was introduced by taking the combination of these two variables. The weights were adjusted such that the new parameter (Average Signal) achieved maximum correlation with both the parameters (correlation of 0.99 and 1 was achieved with signal strength download and signal strength upload respectively) meanwhile also ensuring minimal change in the correlation with the other parameters.

VI. Data Analysis

A. Regression Analysis

It might be a common belief that signal strength and data speed depend on latitude and longitude that is the location of a particular user.

So multiple regression models of degree 1 and degree 2 are built to find how, upload data speed, download data speed, and signal strength depend on latitude and longitude. Here the independent variables are latitude and longitude and hence each regression model has two independent and one dependent variable. To build the regression models normal method was used to obtain coefficients of independent parameters, after minimizing the error function.

The root mean square error also called as RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data that is how close the observed data points are to the model's predicted values. To predict the performance of the models, RMSE is used as the measure. Since all 3 dependent variables did not show much difference in RMSE with models of degree 1 and degree 2, the regression model of degree 1 was used to perform operator wise analysis and eventually test the significance of the models.

For regression of degree 1, significance tests are performed. The F test is used to determine whether a significant relationship exists between the dependent variable and the set of all the independent variables; hence F test is referred to as the test for overall significance. If the F test shows an overall significance, the t-tests are used to determine whether each of the individual independent variables are significant. This

implies that separate t-tests are conducted for each of the independent variables in the model; hence t-tests are called as the test for individual significance. In this paper tests are done 0.05 level of significance. The results are as follows:

- For download data speed and signal strength there exists a significant relationship with latitude and longitude.
- For upload data speed there exists a significant relationship with latitude only.

Hence it can be seen that data speed and signal strength depend on the location. This analysis is further extended to all the operators. Each operator shows slightly different relationships and only RJio reflects the same properties as that of the entire dataset. Comparing the results and the regression models, it is concluded that each operator has different regression equations and slightly different properties but all operators do have a significant relationship between location and data speed and location and signal strength.

This motivates to find a relationship between signal strength and data speed. It is believed that at a particular place high data speed corresponds to high signal strength and vice versa. Hence this can be proved or disproved qualitatively and quantitatively with regression models.

Besides RMSE, another parameter called R-square is also introduced. It is a measure of the goodness of fit of the estimated regression equation. If a variable is added to the model, R-square becomes larger even if the variable added is not statistically significant. The adjusted multiple coefficient of determination compensates for the number of independent variables in the model. In this case also the RMSE value of degree 1 and degree 2 models are similar, therefore the model of degree 1 is used for further analysis. Again the analysis is performed for each operator. Adjusted R-square is also calculated in addition to R-square since there are two independent variables (upload and download data speed). Signal strength serves as the dependent variable.

The results are as follows:

- For the complete dataset that is for all operators as a whole, there exists a significant relationship between signal strength and upload and download data speed.
- The adjusted R-square value is approximately 0.35. This is because varied data of all operators is taken into account.

When the analysis is done for each operator, each operator except Airtel depends on both data speeds. This shows that signal strength depends linearly with data speeds at any location.

All operators except Airtel depend on both data speeds. This shows that signal strength depends linearly with data speeds at any location. The regression equations of Idea and BSNL are similar and hence Idea and BSNL markets can be swapped without much difference in performance (in terms of data speed and signal strength).

B. Clustering Analysis

It has been observed that some areas have proper signal and call quality for particular networks. So the underlying problem statement is that if someone wants to buy a sim of a particular service provider then which provider he /she should go for. Ideally whichever service provider has the best call quality will be the best for that person. Now the call quality of different service providers is not the same throughout the country. Most of the time it varies with region as mentioned above, so it is aimed to suggest the best service provider for a given city to the user.

K means clustering algorithm:

The number of clusters i.e. the K value, the maximum number of epochs, and tolerance levels was fixed. K random points were chosen as the centroid for each cluster and then euclidean distance was used to find the proximity of a point with each of the centroid. The point is assigned to the cluster to which it is the closest. Once the epoch was completed the new means served as the new centroids. The distance from the new centroids was calculated for all the points and the points were again allotted the clusters on the basis of euclidean distance. The difference between the current centroids and the previous centroids was also calculated, this is termed as centroid difference. The maximum number of epochs was fixed as 500 but if the centroid difference comes to be lesser than the tolerance then the algorithm was stopped and the clusters were allotted.

Evaluation of K was done using the elbow method. The sum of squared error is calculated, i.e. sum of squares of the euclidean distance of each point from their respective centroids. An elbow like curve was obtained after which the change in SSE was not that significant. The K at which this occurred was taken as the optimum K. The K chosen for the analysis was 12 in this case. After this the cluster allotment was done using the K means algorithm and the respective clusters were plotted along with their centroids.

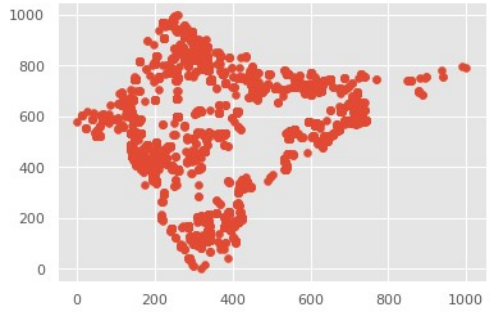

Fig 3: Scatter plot of dataset.

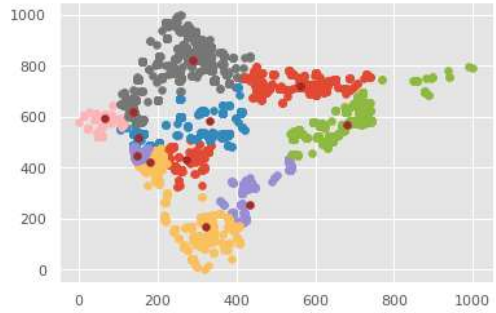

Fig 4: Scatter plot after Clustering.

The geocode API was used to retrieve the latitude and longitude of the city based on the name. The API returned various results out of which latitude and longitude were important. The values obtained were assigned a cluster based on minimum euclidean distance. The mean call rating of each provider was calculated for each cluster. Since the city was allotted a cluster, thus the average ratings for that particular cluster were returned in descending order to give an approximate of the call quality services of the operators in that area. For example, for Mumbai BSNL was found to be the best Network provider in terms of call quality. This would give a fair idea to the customer to choose the best network.

i. Silhouette coefficient

The Silhouette coefficient was also computed and plotted in order to evaluate the algorithm and the K taken. The value of the Silhouette coefficient comes out to be 1 if the point has been allotted to the correct cluster whereas -1 implies that it has been allotted the wrong cluster and 0 implies that it is on the boundary. The mean silhouette coefficient was calculated and it came out to be approximately 0.73 which is close to 1. The following plot was obtained for the silhouette coefficient for each point in their respective allotted cluster.

C. Logistic Regression

The relation of call satisfaction was found with other parameters using Logistic Regression. Data had to undergo some processing. In the data call satisfaction had 3 levels, namely 'Satisfactory', 'Call Dropped' and 'Poor Voice Quality'. For the analysis it was sufficient to know if the call was satisfactory or not, so a new parameter was created named 'call satisfaction' containing values 0 and 1. Data entries with category 'Satisfactory' in Call Drop Category were coded as 1 while and data entries with categories 'Call Dropped' or 'Poor Voice Quality' were coded as 0. After encoding the data had 70 %(approx.) Values as 1 and only 30% of the values as 0. Thus down sampling was performed for the data before training the models. There were 4 models implemented namely, call satisfaction vs latitude and longitude, call satisfaction vs data speed, call satisfaction vs data speed download and signal strength, call satisfaction vs data speed upload and signal strength. The first model gave an accuracy of approximately 70%.

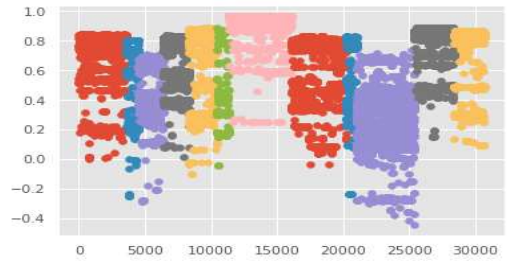

Fig 5:

This graph shows that most of the values are near 1 thus the points have been allotted to correct clusters.

D. Chi-square tests

A chi-square statistic is a test that measures how expectations compare to actual observed data. An important application of a chi-square test involves using sample data to test for the independence of two categorical variables. The null hypothesis for this test is that the two categorical variables are independent. Thus, the test is referred to as a test of independence. The steps for chi-square tests are: State the null and alternative hypotheses. Record the observed frequencies, f_{ij}, in a table with r rows and c columns. Assume the null hypothesis is true and compute the expected frequencies, e_{ij}. Compute the chi-square statistic and then apply the rejection rule. Chi-square distribution has (rows – 1)(columns – 1) degrees of freedom and is the level of significance for the test.
In the data Population, Literacy Rate and Rainfall were continuous features. After the features were normalized, they were discretized with equal interval binning. Then three chi-square tests were performed: Literacy Rate vs Network Type, Rainfall vs rating and Population vs rating. For each test, $\chi^2 \geq \chi^2_\alpha$, hence the null hypothesis was rejected and so it could be concluded that the two categorical variables in each set are not independent.

VII. Results and Discussion

The data set analysis led to some insightful results. To mention a few: data speed and signal strength were linearly related for each operator and for the dataset as a whole. BSNL and Idea were similar in terms of speed and signal strength and hence their markets could be swapped without much change in performance. Data speed and signal strengths varied with location linearly. K Mean clustering helped to predict the best operator for a particular city based on the ratings. The data analysis techniques provided a holistic view of the features and their relations with each other. This analysis could be extended to bigger and uniform datasets for each month to predict the seasonal variations and get concrete results.

VIII. Acknowledgment

We would like to thank our Data Mining Professor Dr. Manik Gupta. We would also like to thank the TAs who helped us in choosing the dataset and gave us valuable inputs. We would also like to thank BITS PILANI Hyderabad Campus for providing such an opportunity to work upon this.

Reference

[1] Tan, P.-N., Steinbach, M., and Kumar, V. 2005. Introduction to Data Mining. Addison-Wesley.

[2] https://searchbusinessanalytics.techtarget.com/definition/data-visualization

[3] http://www.cs.put.poznan.pl/jstefanowski/sed/DM14-visualisation.pdf

[4] https://www.techopedia.com/definition/14650/data-preprocessing

[5] https://medium.com/@saishruthi.tn/data-mining-introduction-data-preprocessing-5080be604f96

[6] https://docs.microsoft.com/en-us/analysis-services/data-mining/discretization-methods-data-mining?view=asallproducts-allversions

[7] https://towardsdatascience.com/data-visualization-and-its-techniques-454ab3f31bf7

[8] Github repository link: https://github.com/DrishtiMamtani/Call-Quality-Analysis