NETWORK PROVIDER RECOMMENDER SYSTEM

(Data Mining Project Presentation (CS F415))

TEAM 16 :

Aniruddh Gupta (2017A7PS0149H)

Vidish Bharadwaj(2017A4PS1391H)

Anuj Kharabanda(2017B4A71508H)

Anushray Mathur(2017A7PS1570H)

CONTENTS

PROBLEM OVERVIEW:

It is a widely known fact that cellular communication companies differ from each other considerably in terms of **voice call quality** when pitted against one another under the same circumstances. This difference can arise due to several reasons like **Network Blind Spots, low bandwidth allocation** due to loaded cell sites, **sparsely located network towers** etc. To the average person selecting a network carrier, this is a problem too complex to be solved manually, hence the need for a recommender system that takes into account various external factors and gives the most reliable network provider for the given case as output. System will recommend the **best possible network provider** by seeking requirements from the user such as their location, network type they want to use and use type, predict the **expected call drop quality** based on network operator, use type, network type and location coordinates. We will also find relevant **association rules** between the attributes of the data set.

DATA OVERVIEW:

This data set was recorded monthly at various locations throughout India consisting of around 25,000 entries (per month) with the following attributes:

1. Operator (Nominal)
2. Use Type (In/Out/Travelling) (Nominal)
3. Network Type (2G/3G/4G) (Nominal)
4. Call Quality Rating (out of 5) (Ordinal)
5. Call Drop Category (Ordinal)
6. Latitude (Ratio)
7. Longitude (Ratio)
8. State Name (Nominal)

Our dataset includes the data for the months of April 2018 to March 2019.

DATA PRE-PROCESSING:

To make the dataset more efficient to use and get it in a useful format, the following operations were performed:

1. **Data Cleaning**
2. **Data Transformation**
3. **Data Visualisation**
4. **Exploratory Data Analysis**

The aforementioned operations are elaborated upon in the subsequent slides.

DATA CLEANING

- Handling Missing Values: Added the state value if valid location coordinates for that entry was available

- Removed entries with irrelevant values(operator, type and location coordinates having undefined values)

DATA TRANSFORMATION

- Label Encoding was used on this attribute to transform the values from **Call Dropped, Poor Voice Quality and Satisfactory** to **0, 1 and 2** respectively. This was done to create an ordinal relationship between the values of the attribute and make the later operations easier to perform.

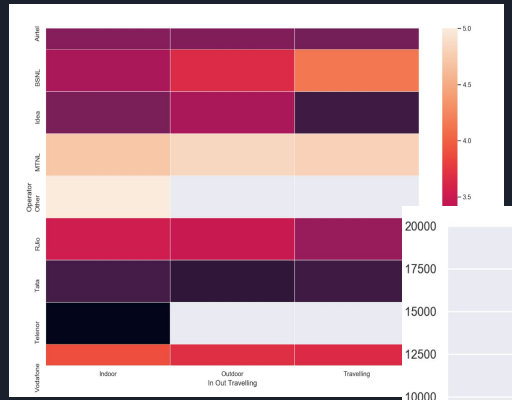
EXPLORATORY DATA ANALYSIS:

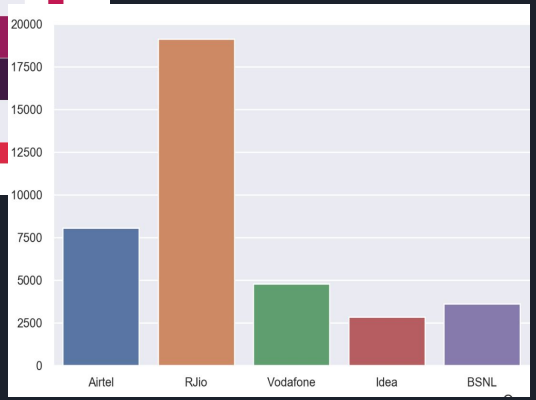
- **Correlation:**
  - **Pearson's**: Pearson's correlation coefficient is a statistical measure of the strength of a **linear** relationship between paired data.
  - **Spearman's**: Spearman's correlation coefficient is a statistical measure of the strength of a **monotonic** relationship between paired data.
- **Chi Square Test:**
  - To find dependency between rating and call drop category

VISUALIZATION:

The following plots were created during the course of data visualization of the dataset for each of the 12 months' data as well as for the combined data:

- Distribution Plots
- Bar Plots
- Box Plots
- Scatter Plots
- Heat Maps
- Correlation:





DATA PROCESSING TECHNIQUES:

The chosen data processing techniques were applied on the pre-processed data set to get insightful results and statistical inferences about the data provided. These techniques were also deemed necessary to be able to predict the end user experience based on given attribute data for any new data entry.

The applied Data Processing Techniques were as follows :

- **K.N.N. Classification**
- **Association Rule Mining**
- **Decision Tree Classification**

The process details and results of the aforementioned techniques are presented in the subsequent parts.

K.N.N. CLASSIFICATION:

The K.N.N. Classification model developed uses the following attributes as input to predict the call drop category:

- **Operator**
- **In/Out/Travelling**
- **Network Type (2G/3G/4G)**
- **Location Coordinates**

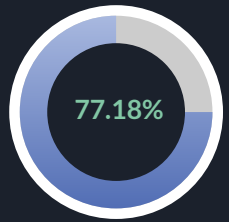
First the model filters the data set based on the given input and then applies K.N.N. Classification on the data set to predict the call drop category.

REASON FOR THE CHOICE OF K.N.N. CLASSIFICATION OVER CLUSTERING :

Clustering was not deemed feasible as when we plotted the training data set on the 2-D plane, no meaningful sets of clusters were visible. Hence, the pursuit of Clustering was dropped in favour of going towards the K.N.N. Classification route.

RESULTS OF K.N.N. CLASSIFICATION :

Optimal **K** value which was used in the model was = **101**.
It Gave **82.03%** accuracy with **128** testing entries.
Accuracy with which the model predicts the target class is **77.18±1.59 %** .



ASSOCIATION RULE MINING :

Apriori Algorithm was implemented in order to find the Association rules for the data sets given to us. The data sets were concatenated into a single large data set for ease of use and availability of sufficient number of rule occurrences.

The dataset was one-hot encoded at first, after which the frequent itemsets were found. Consequently, the required rules were formulated using different combinations of minimum **support,** minimum **confidence** and **lift** so as to find the best combinations which provided the required results.

RESULTS OF ASSOCIATION RULE MINING :

The minimum support used in the model was **0.04** while the value of minimum confidence and minimum lift are **0.8** and **1.9** respectively. Using this combination the following **12 rules** were mined (single digit numbers denote Call Rating):

- **['Call Dropped'] -> [1]**
- **[2] -> ['Poor Voice Quality']**
- **['4G', 'Call Dropped'] -> [1]**
- **['Indoor', 'Call Dropped'] -> [1]**
- **['4G', 2] -> ['Poor Voice Quality']**
- **['Indoor', 2] -> ['Poor Voice Quality']**
- **[4, 'RJio'] -> ['4G', 'Satisfactory']**
- **[5, 'RJio'] -> ['4G', 'Satisfactory']**
- **['RJio', 4, 'Maharashtra'] -> ['4G', 'Satisfactory']**
- **['Indoor', 4, 'RJio'] -> ['4G', 'Satisfactory']**
- **[5, 'Outdoor', 'RJio'] -> ['4G', 'Satisfactory']**
- **['Indoor', 5, 'RJio'] -> ['4G', 'Satisfactory']**

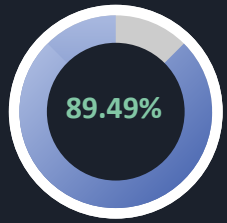
DECISION TREE AND RANDOM FOREST CLASSIFICATION:

Given all the details as input from the user, our aim was to find the best network operator for the user. This model accomplishes the same using implementations of **Decision Tree** and **Random Forest** **made out of many decision trees.**

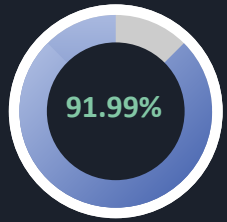
1. The **Decision Tree** algorithm creates a decision tree with all features (except operator) as parameters and predicts the operator.
2. For **Random Forest**, we create a set of decision trees with bootstrapped data, i.e., only a percentage of data from the whole dataset and random subspace of some features among all the features

RESULTS OF DECISION TREE AND RANDOM FOREST CLASSIFICATION :

- The **decision tree** algorithm has an average accuracy of 0.906437 **(90.64%)** and a total accuracy (after appending all data) is 0.89493 **(89.49%)**.



- The **random forest** algorithm has an average accuracy of 0.919886 **(91.99%)**.



The total accuracy of the random forest wasn't calculated because of computational constraint as it took nearly a day to predict total accuracy of the decision tree and hours to predict accuracy of even one month in the random forest.