**Abstract**

- Are you happy with the quality of service you get from your wireless provider?
- How does quality relate with geographic location or time of day?
- Do you know that your service provider does talk about the coverage in an area but does not guarantee any quality of service.
- Can you compare your service provider with others?

Research done by Federal Communications Commission (1) enable us to compare coverage provided by different providers but still does not comment on the quality of service. This project aims to make a statement about the quality of service provided by different service providers depending on various aspects like geo location, network type, providers depending on various aspects like time, geographic area and technology.

**Objective**

- Use big data analysis to:
  - Find correlation amongst various features.
  - Improve the bandwidth testing methodology used previously.
  - Determine the quality of data from android application.

**Methodology Steps**

<table>
<thead>
<tr>
<th>Step No.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Data Cleaning and Preparation (Missing data, Class imbalance, anomalies, outliers) (R, Python JSON Flattening)</td>
</tr>
<tr>
<td>2</td>
<td>Exploratory Data analysis (Histograms, binning, scatter plots)</td>
</tr>
<tr>
<td>3</td>
<td>Attribute Selection (Random Forest, Bag/Boost, PCA)</td>
</tr>
<tr>
<td>4</td>
<td>Finding Correlations (Spearman, Pearson, Linear, Logistic Regression)</td>
</tr>
<tr>
<td>5</td>
<td>Data mining algorithms implementations (J48, Random Tree, Random Forest)</td>
</tr>
</tbody>
</table>

**Data Cleaning Process**

- Missing values in the dataset are populated using Multiple Imputation Chained Equations package (MICE) in R. In this method, missing value for each attribute is calculated by regressing the attribute against all other attributes.
  - Posterior Predictive Distribution: Variation of the Bayesian distribution.
  - Regression Algorithm
  - Stabilizing the imputation results obtained.
- Class Imbalance: Synthetic Minority Oversampling Technique (SMOTE)
- Preprocessing the nested JSON files for bad data. As shown in the figure above, we store the bad data from the android application in the form of JSON documents. This JSON document contains the hierarchy of all the JSON documents involved within the same call or bandwidth test.
- Reverse Geocoding using Python. In order to relate the base stations and geographical coordinates we used python's geocoding package.
- Feature Selection: Done using Principle Component Analysis (PCA) and Random Forest (Ensemble method).

**Results**

- Figure 6 shows the comparison of quality of service provided by various radio technologies.
- Figure 7 shows the decision tree obtained after we ran REPTree and Random forest algorithms on the data collected.
- REPTree: Reduced Error Pruning
- Random forest: Ensemble method combining weak methods into one strong method.

**Conclusions**

- Our results show that analysis of bad (erroneous) data can provide us a good pattern and helps us understand the problems in the software development cycle better.
- We were not only able to classify the good data from bad data but also provide the pattern about why the instances might be getting rejected.
- Understanding the data collection process helps us to tackle the problem at hand better.
- When doing data analysis, always make use of more than one methods to cross verify the results obtained from the previous methods.

**Future Work**

- This project lays a foundation of the system where consumers can make an informed decision about which service provider is better based on consumer needs.
- This can also be used to recognize and handle the problems with the software development lifecycle.
- This can be useful to the service providers to keep a track of the quality of the service based on time, geographic area, technology.

**References**


---

**Acknowledgement**

Professor Dr. Yossi Nygute
College Of Applied Science and Technology
Professor Dr. Clark Hochgraf
Associate Professor of Electrical, Computer, and Telecommunications Engineering Technology