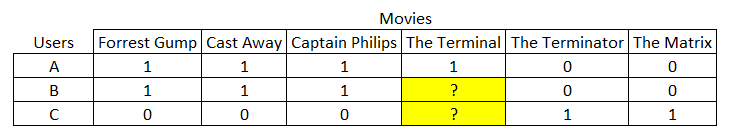
**Recommender systems - theory**

Similarity based collaborative filtering:

Required: User + item (movie/product) interaction data only

Ex: User- movie matrix



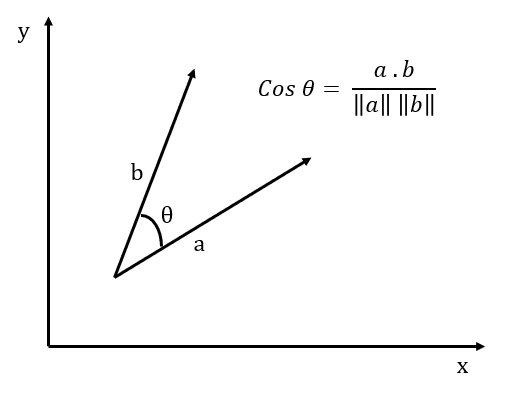
Creating an interaction matrix with pivot in panda:

***interactions\_matrix = movies.pivot(index="userId", columns="movieId", values="rating")***

Fill missing values with 0:

***df.fillna(0)***

**Similarity based recommendations: Concept**



**Common details**

Research Aims:

* Introduce an optimized hybrid recommender systems for telcom industries (eg. Sri Lanka Telecom) to recommend telcom services to a targeted customers/ subscribers, which can add a business value in
  + Increasing revenue and growth
  + Reducing customer/subscriber churn
  + Empowering long-term customer relationships
* Telcom industries have highly asynchronous and significantly different user-service/interactions dataset depending on a unique set of variables compared to other fields, as retail, movies etc.. where a lot of recommender systems have been already implemented. So there is a very few attempts made to propose a recommender system for telcom domain.
* Most of the proposed recommender systems are content-based or similarity based, where it suggests a service to a subscriber either rule based or depending on the cosine similarities [] / correlations like statistical approaches.
* Customer/subscriber segmentation in required, prior providing any sort of recommendation, subscriber base should be clearly identified and taken into account when recommending new services/ offers in order to reach business aims such as increasing profitability and reducing subscriber churn.
* There is no study present, suggested to analyze subscriber reviews and consider into the content for the persistent recommendation model, when recommending new services. If more customer staisfied services (services with a good sentiment among the subscriber base) can be recommended frequently, more profits/ benefits can be gained.

Methodology/ Technology:

* The study investigates RS approaches that contain ML algorithms and that were implemented and tested with real or simulation data
* Baseline Study Methodology (data gathering): data dump/ row data collected from a telcom giant in Industry (SLT)
* Scraping reviews, comments etc..
* Subscriber Survey : An independent survey conducted among stratified unbiased sample of the telcom users (population) in sri lanka
* <http://www.trc.gov.lk/> - The Telecommunications Regulatory Commission of Sri Lanka (TRCSL)
* Design of Software Project:
  + Examine current recommender systems and their approaches/ algorithms

Ex: Amazon, eBay, Facebook etc..

* Compare and contrast relevant techniques/ methodologies for telcom customer segmentation (market basket analysis/ uplift modeling etc.)
* Compare and contrast relevant algorithms/ methodologies for recommender systems
  + Content based
  + Collaborative Filtering (Matrix Factorization etc.)
  + Similarity based
  + ALS
  + Random walks (Network analysis)
  + Association Rule Mining (Matrix)
  + Mean Average Precision (MAP)
* Express research Gap/ uniqueness of telcom domain, and requirement of the proposed hybrid model
* Support of customer review/rating analysis for finding/predicting user-product interactions (as an input to the hybrid recommender algorithm for better results/accurate predictions)
* Develop a web based application with Graphical User Interface, to integrated recommendations, and a dashboard (kibana/grafana or streamlit) to visualize insights and findings)

Techniques to measure validity/ accuracy:

* To compare algorithms and their performance : use of a confusion matrix
  + A confusion matrix illustrates the accuracy of the solution to a classification problem.Given n classes a confusion matrix is a m x n matrix, where Ci,j indicates the number of tuples from D that were assign to class Ci,j but where the correct class is Ci . Obviously, the best solution will have only zero values outside the diagonal a confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. Some standards and terms:

1. True positive (TP): If the outcome from a prediction is p and the actual value is also p, then it is called a true positive.

2. False positive (FP): However, if the actual value is n then it is said to be a false positive.

3. Precision and recall: Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

Telecom Industry do possess huge amounts of information, maybe more than any

organization in the world . In any case, the battle lies in catching the insights from this

unstructured datset and drawing business insights using the knowledge of ML

Collaborative filtering: user based/ item based

content-based recommender systems do not use other user’s ratings at all. Instead, they utilize descriptive keywords associated with each item to make recommendations.

Recs we have:

Rule Based

NN

Basyen

Decision Tree

Random Forest

Statistical Model - ML based

Deep Learning

Content based : ALS/ Apropri

Matrix Factorization

Coefficients/ Correlations

Collaborative (ex: Cosine similarities)

What we build: (Research Problem)

For telecom industry

* Find right customer base (uplift-modeling based customer segmentation)
* Provide telecom service recommendations - Content based/ collaborative hybrid model

Steps:

* Dataset gathering
* Data engineering
* Dimension reduction
* Subscriber segmentation with uplift modeling
* Comparing complexity/ performance/ accuracy on algo for a sample set (cofussion matrix etc)
* Provide recommendations algo (hybrid)
* Evaluate/ Validation

Research Gap: (Whats New)

* Only product/service recommenders (only a few has methodologies, other are review papers and suggestions)
* Telcom domain has significantly different data, other than generally available or used in most commonly researched RSs.
  + common: User demographics/ user interactions/ user reviews
  + Difference:
    - to recommend: services not products / service consuming time / frequency/ location based differences / mobile device based differences/
    - To segment: value added/loyalty customers / usage history/ buying patterns/
    - Asynchronous behavior/ real time recs neededRecommender service is also able to receive a synchronous event in order to set
    - **Realtime recommendation (using cloud tech)**
    - **Review based recommendation**
    - the algorithm to be used for the prediction:

Dataset: Collected from SLT (from their data warehouse)

* No of rows/ cols/ details of the cols

**IMPORTANT**

\*Conducting a Algorithm Comparison: Confusion Matrix

\*Experimenting Performance: (For different sizes of test sets) - Ex: 1000/5000/10000 rows/ no of iterations - check for complexity

***Scihub***

**Coefficients:**

**Pearson**

**Spearman  
Correlations:**

**Cosine**

**In telcom services:**

acquiring new customers is multiple times more expensive than retaining a customer,

operators are willing to pay for software that can help them identify customers with high revenue contributions but are likely to terminate the service. (paper 04)

traditionally, statistical methods such as decision tree are used to classify customers

into buckets with varying levels of churning tendency (e.g.,low, middle, and high) or probability.

Subscribers are more likely than before to churn because the inconvenience of having to change to a new phone number no longer exists;

**Note:**  (paper 04)

statistical methods such as decision trees have long been used to predict subscriber churn likelihood.

Usually these methods look at subscriber demographic data such as age, occupation, gender,

income and call activities from CDRs to build the predictive model. Statistical models are excellent at quantifying and correlating these attributes.   
  
**They however fall short in causation analysis, i.e. they do not tell you what the causes**

**of the churns are. Are people leaving because the quality of the network is bad or subscriber specific reason. (is the campaign/ recommendations right?)**

With the advancement and availability of big data technologies, operators are now able to collect more nearly complete data about a user’s experience and behavior. They

can then build a behavior-based churn prediction model to alert them when a subscriber is about to leave.

**Text analytics can also be applied to CRM logs and/or social media to gauge customer sentiment and identify those that may leave**

Accurate measurement of customer QoE can help operators to predict customer churns and identify customers for service upgrade or target marketing.

**The data may include, for example, user locations, click stream logs, app usage, network performance, subscriber plans, and demography,and call center logs. Each user session (e.g., video, voice, or Web sessions) is associated with a set of service KPIs (S-KPIs) calculated based on the raw network data.**

**The major focus of most subscription-based companies is to acquire new customers and to maintain old customers by identifying customers with a major propensity to churn.**

**Thus, Customer churn has only results in opportunity costs due to reduced sales, but also causes negative word-of-mouth opinions about the organization, which can be detrimental to the company's reputation.**

The interest of a user on an item is usually measured by a rating which can be obtained either

explicitly or implicitly.

* Numeric: when numerical values are assigned to products/services, for example, the five stars on the Amazon website
* Ordinal: when the user is prompted to select a term that best indicates his/her opinion on an item, such as “I agree”, “I am neutral” and “I disagree”
* Binary: when the user simply decides if an item is good or bad
* Unary: this kind of ratings was popularized by Facebook where users can mark his/her interest in a post or photo by clicking a button “Like”

there are two main methods of **collaborative filtering** , the nearest neighbor methods and the latent factor methods.

The nearest neighbor methods are based on the principle that users who have preferred similar items in the past tend to prefer similar items in the future. These methods can be user-based or item-based. In the user-based collaborative filtering, the items (content, services, products, etc.) recommended to a user are those that other users, with similar preferences, have chosen previously. User-based collaborative methods firstly find the users more close to each user, i.e., those with more similar taste and preference. Then, only items that are preferred by these users are recommended to the target user.

In the item-based collaborative methods the similarities among different items in the

dataset are calculated by using a similarity measure, and then these similarity values are used

to predict ratings for user–item pairs not present in the data. The latent factor methods, in turn,

are intended to explain users’ preferences characterizing users and items by factors, which are

characteristics and patterns inferred from existing assessment data.

Collaborative filtering: present a problem known as “cold-start”. This problem occurs when the system can not make reliable recommendations due to the lack of initial ratings or necessary information.

**Content-based filtering:**

the items recommended to a user have similar content to the items that this user chose in the past, that is, only the items of high similarity with past user preferences are recommended. Content-based filtering methods have the advantage of not being dependent on the ratings of other users.

Problems: limited analysis of the contents, which refers to the difficulty in extracting reliable information automatically from various content such as images, video, audio and text; and (ii) super-specialization, as the system recommends items analyzing the user profile, the user is restricted to see similar items to those that he/she has already assessed/seen before.

**Context-aware Filtering:**

**According to Adomavicius and Tuzhilin**

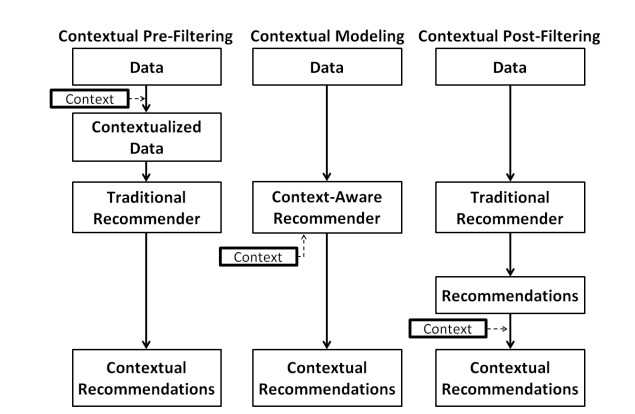
Adomavicius, G.; Tuzhilin, A. Context-Aware Recommender Systems. In Recommender Systems Handbook;

Springer: Berlin/Heidelberg, Germany, 2011; pp. 217–253.

the systems can be divided into three categories

(i) contextual pre-filter,

(ii) contextual modeling, and

(iii) contextual post-filtering.

* Need to filter outliers

Opinion Mining Intro:

With the growth of social networks, more and more users can openly discuss their impressions

and experiences on a variety of products, items, and services. This means a significant increase in user-generated content in the form of reviews, blogs, discussion forums, social networks, etc. Among this content, reviews represent rich sources of data and they are very useful for marketing intelligence, social psychology and other areas that are interested in mining opinions, views, sentiments and attitudes

Analysing sentiment of reviews:

Frequency-based: an aspect can be expressed by a noun, adjective, verb or adverb, but studies

show that from 60 to 70% of explicit aspects are nouns [69]. Aspects tend to be frequent nouns

since, in commentaries, people are generally more likely to talk about the relevant aspects.

Based on syntactic relations: there are usually many syntactic relationships between the

expressions of sentiment and the opinion targets. Such relationships are possible to be explored

when words and phrases of sentiment are known.

Through supervised learning: in general, methods for identifying aspects are based on sequential labeling. The most commonly used methods are: **Conditional Random Field (CRF)** [74] and **Hidden Markov Model (HMM)** [75].

Uplift Modeling:

* Basically calculating the probability of a potential lead being converted into a positive impact by some sort of treatment such as a marketing offer (models the incremental impact of the treatment). so that focus can be given to the customers with higher potential to be converted by the treatment.

### Based on customer behaviour

* **Sure things**: People who will purchase no matter what
* **Persuadables**: People who will purchase only if they are exposed to a recommendation
* **Lost causes**: People who will not purchase no matter what
* **Do not Disturb / Sleeping dogs**: People who will not purchase if they are exposed to a recommendation



* Further to propensity calculation, we have to separate out the customers who were going to buy anyway from the ones that need to be persuaded.
* Points to be considered,
  + “Did my recommendation cause the customer to purchase from me?”
  + “Did I waste money recommending to customers who were already going to purchase from me?”
  + Did my recommendation make the probability of someone purchasing worse (negative impact)?

Importance of Uplift modeling:

* **Classic propensity model (and most machine learning algorithms) predict target (y) given features (x). Uplift looks to solve for the impact of treatment (t) on target (y) given features (x).**
* A propensity model adds value by helping you avoid lost causes. Uplift modeling improves the targeting even further by focusing on only the customers who are in the persuadables segment.

## Direct Uplift Models

* Requires the algorithm like random forest to be completely redesigned to feature select, hyper parameter tune, and fit to solve for the impact of (t) on (y) given (x).
* Limited to Forests mostly. No XGB or NN.
* You can implement uplift trees with the package Uplift for R and CausalML for Python
* Likely to produce most accurate results
* Tend to be slower

## Meta Learners

* Transform the problem(e.g. to a set of regression problems) and use existing algorithms to indirectly estimate the uplift.

Evaluate the modeL:

## By Decile

* We can not have actual uplift value for a given customer, since he can not be both given and not given the treatment. In that case we group a set of customers and calculate actual uplift for that entire set.
* In order to divide customers into groups and to have the customers of the same segment (based on customer behaviour), we order them by order of predicted uplift scores. In such cases persuadables would have the higher values, sure things / lost cases would have values near to 0 and DNB ones should have a negative value.
* Say we divide the predicted scores into deciles and calculate actual uplift scores for the customers as a whole in each decile. With this setting, the ideal graph should be like the following (One in Red).

# Challenges

* Ground Truth not available
  + This will only impact the evaluation stage.
  + A person can not both have and don’t have the treatment
  + Making the ground truth an estimated value, impacting the model accuracy calculations.
* Limited no of packages available

**Research Paper 01**

* Recommender System for Telecommunication Industries:

A Case of Zambia Telecoms

* American Journal of Economics 2017, 7(6): 271-273
* DOI: 10.5923/j.economics.20170706.01
* Mulizwa Soft 1,\* , David Makadani Zulu 1 , Ruzive Mazhandu 2

Recommender systems use machine learning algorithms and artificial intelligence techniques to recommend products to customers. These algorithms use historical data of purchases of other people to determine which products to recommend to a particular customer, in general recommender systems are designed in such a way that they automatically generate personalized suggestions of products to customers  
  
Recommender based systems implemented using the concept of big data, machine learning algorithms or deep learning algorithms have a lot of advantages which benefits a lot of companies, for example better user experience, increased average order value, Increased Sales and improved customer retention.

Recommender systems use the concept of rating to measure user’s preferences and a range of filtering techniques, and can be classified in multiple ways according to the nature of the input information. The content-based (CB) methods and collaborative filtering (CF) methods are the most popular techniques adopted in recommender systems. The CB methods recommend products by comparing the content or profile of the unknown products to those products that are preferred by the target user. However, these methods tend to rely heavily on textual descriptions of items, leading to several unsolved problems such as limited information retrieval, new user

problems, and overspecialization.

Problems:

* lose their revenues because of customers switching from one provider to another in search of cheap affordable high-quality products and services. This causes subscriber base and revenue shrinkage, the increase in churn rate causes a loss of future incomes
* When the right recommendations were not provided to the right customer (loss due to recent disconnects/ loyalty benefits provision)
* When lack of a proper understanding of the both current and future(predictive) customer base, their (predictive) requirements and which services/ products/ packages to offer accordingly in future

Research Questions:

* How do we analyze the relationship between telecommunication subscribers and

telecommunication products?

* What are the challenges telecommunication companies face that lead to high revenue loss, churn and bad customer experience?
* What are the identifiable customer segments and how to determine which services/ packages are most applicable/ profitable targeting these segmented groups
* How to process high volumed big data, frequently to capture those insights?

**Research Paper 02**

The Use of Machine Learning Algorithms in Recommender Systems: A Systematic Review

* Ivens Portugal Paulo Alencar Donald Cowan
* November 2015 Expert Systems with Applications 97
* [Submitted on 17 Nov 2015 (v1), last revised 24 Feb 2016 (this version, v4)
* Cite as: arXiv:1511.05263 [cs.SE]

However, choosing a suitable machine learning algorithm for a recommender system is difficult because of the number of algorithms described in the literature. Researchers and practitioners developing recommender systems are left with little information about the current approaches in algorithm usage. Moreover, the development of a recommender system using a machine learning algorithm often has problems and open questions that must be evaluated, so software engineers know where to focus research efforts.

The field of RS has its origins in the mid-1990s with the introduction of Tapestry [23], the first RS. As the RS field evolved, researchers studied the use of algorithms from machine learning (ML), an area of artificial intelligence (AI). Machine learning has been studied since the late 1950s [40], with the emergence of the field of AI.

Today, there is a plethora of ML algorithms (k-nearest neighbor [48], clustering [28], Bayes network [22], to name a few types), which are used in applications of a huge range is vast and the field looks very promising.

However, the ML field does not have a clear classification scheme for its algorithms, mainly because of the number of approaches and the variations proposed in the literature [37]. As a consequence, it becomes difficult and confusing to choose an ML algorithm that fits one’s need

For these reasons, it was decided to do a systematic review to investigate how real implemented RSs with ML algorithms are being used; which ML algorithm they use; and how the SE field can impact their development. when developing an RS. In addition, researchers may find it challenging to track the use of ML algorithms in RSs.

**IMPORTANT**

RSs using a collaborative approach consider the user data when processing information

for recommendation. For instance, by accessing user profiles in an online music store, the

RS has access to all the user data, such as the age, country, city, and songs purchased.

With this information, the system can identify users that share the same music preference,

and then suggest songs bought by similar users.

RSs with a content-based filtering approach base their recommendations on the item data

they can access. As an example, consider a user who is looking for a new computer using

an online store. When the user browses a particular computer (item), the RS gathers

information about that computer and searchers in a database for computers that have

similar attributes, such as price, CPU speed, and memory capacity. The result of this

search is then returned to the user as recommendations.

The third classification describes RSs that combine the two previous classifications into a

hybrid filtering approach, recommending items based on the user and the item data. For

example, on a social network, an RS may recommend profiles that are similar to the user

(collaborative filtering), by comparing their interests. In a second step, the system may

consider the recommended profiles as items and thus access their data to search for new

similar profiles (content-based filtering). In the end, both sets of profiles are returned as

Recommendations.

Collaborative Filtering - Data Gathering:

* Explicit (User is aware that they are providing information) - General info and their propensity/ likelihood ex: using a likert scale
* Implicit - User is not aware/ there data collected on actions (Product Views/ clicks/ car adds etc..) (other than personal data given when registering)

Context-aware recommendations [3] are based on the context of the user. A context is a set of information about the current state of the user, such as the time at the user location (morning, afternoon, evening), or their activity (idle, running, sleeping). The amount of context information to be processed is high, making context-aware recommendations a challenging research field

**ML Intro:**

Learning is the process of knowledge acquisition. Humans naturally learn from

experience because of their ability to reason. In contrast, computers do not learn by

reasoning, but learn with algorithms.

In unsupervised learning, ML algorithms do not have a training set. They are presented

with some data about the real world and have to learn from that data on their own.

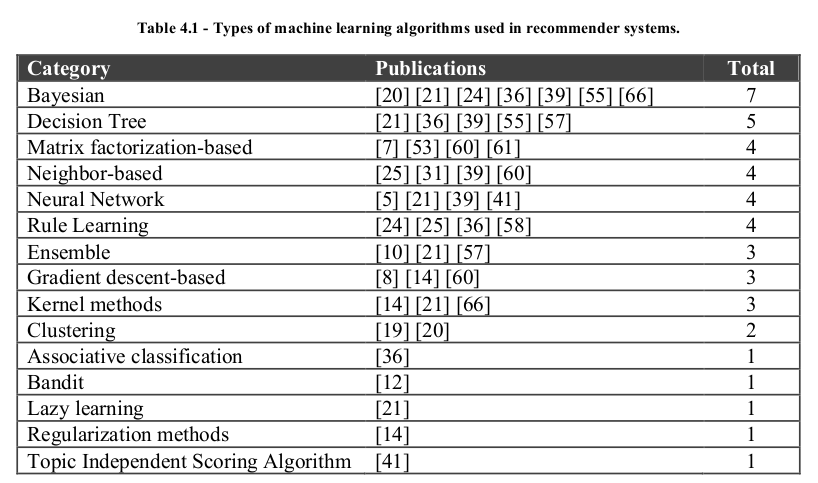
Unsupervised learning algorithms are mostly focused on finding hidden patterns in data.

For example, suppose that an ML algorithm has access to user profile information in a

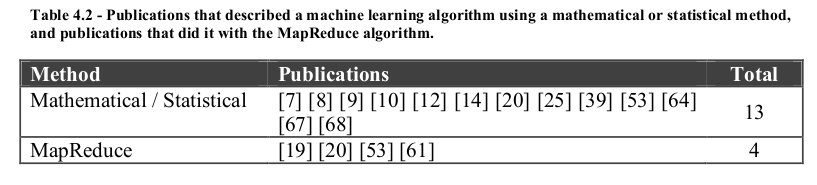
social network. By using an unsupervised learning approach, the algorithm can separate

users into personality categories, such as outgoing and reserved, allowing the social

network company to target advertising more directly at specific groups of users.



The same explanation may justify the use of Decision Trees as the second most used approach for ML algorithms in RS. Both Bayesian approaches and Decision Trees have similar underlying calculations and recently have become popular.



Total : 26 publications

Mathematical or statistical methods are calculations such as cosine measure [18], least

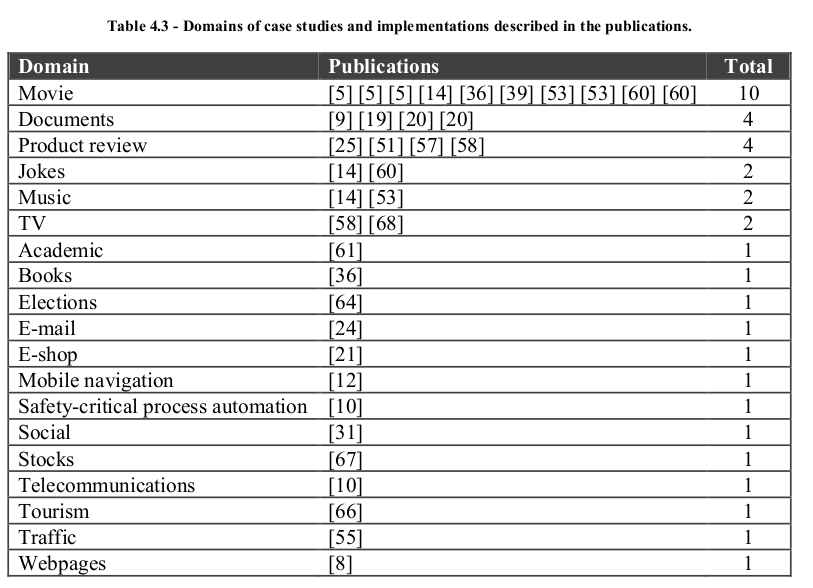
squares [1] or Pearson correlation [18]

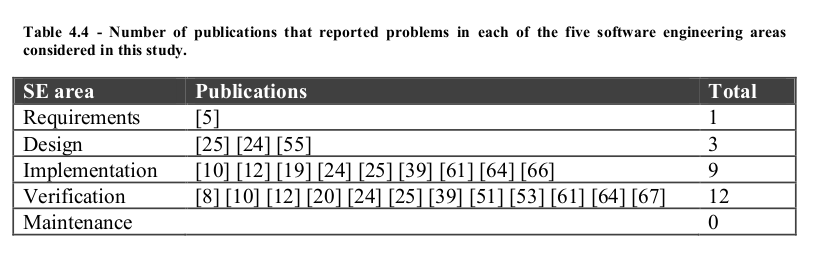
The MapReduce algorithm [17], which became popular in mid-1990s, is a programming model suited for distributed and parallel computation.

**Areas of Research in RS Papers:**

The domain of Movies is used the most with 10 occurrences in the 26 publications. One

reason for this result is the ease of access to data in the movie domain.





**RESEARCH PAPER 03**

A recommender System for Telcom Users

* Software Engineering Research Group, Dept. of Control and Computer Engineering,
* Politecnico di Torino, 10129, Torino, Italy
* Paolo Falcarin, Antonio Vetrò, Jian Yu

**IMPORTANT: Intro:**

The increasing flourish of available services in telecom domain

offers more choices to the end user. On the other hand, such wide offer cannot

be completely evaluated by the user, and some services may pass unobserved

even if useful. To face this issue, the usage of recommendation systems in

telecom domain is growing,

Recommendation can be seen as an advanced form of personalization, because user preferences are used to predict the interests of users for a new service.

Furthermore, telecom services now can be created and provisioned by end users [1],

which will result in more services being available. Given this background, one serious

question that needs to be answered is how to avoid telecom users getting lost with the

huge amount of available services.

In fact, the wide availability of users’ data in the telecom domain is a good starting point for applying different analysis in order to suggest the more suitable services to a target

User.

and recently there are some works on service recommendation [6][7]. They made good foundation for our work, but characteristics of telecom domain should be tackled to apply the recommendation in telecom domain successfully

In telecom domain, the related work on recommendation is still not much. Ricci et

al. [6] discussed a kind of mobile recommender system. Chen et al [7] present a

recommendation algorithm in mobile environment, but they do not mention the

corresponding architecture of the recommender system.

The main characteristics of our work is the evaluation of different correlation

algorithms on a typical telecom data set,

**IMPORTANT**

There are some specifications of telecom service platform, born to combine

telecom resources and IT systems, such as Java Service Logic Execution Environment

(JAIN SLEE) [4] and Parlay [5]. JAIN SLEE is chosen in our work for its event-

based architecture and the easiness of integrating IT services. A Service Logic

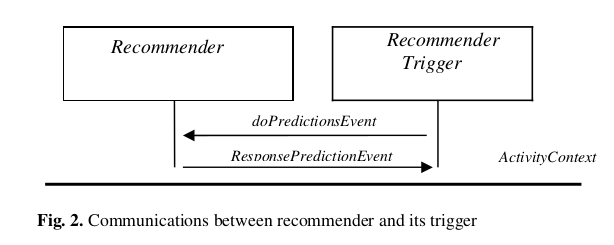
Execution Environment (SLEE) is a high throughput, low latency event processing

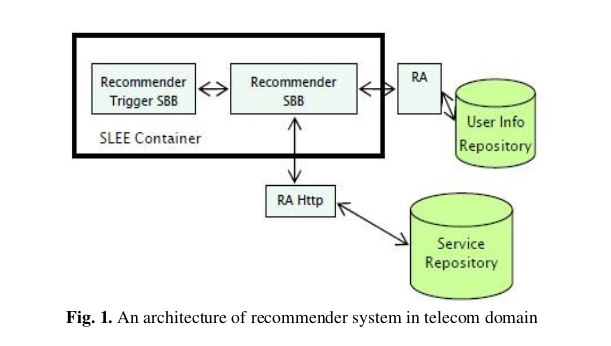
application environment used in telecommunications industry.

Unlike enterprise applications which are usually invoked synchronously, telecom

applications are always invoked asynchronously.

Real Time Recs Prediction for Telcom: Architecture





**IMPORTANT**

The methodology this paper suggested:

the n highest recommendations of

services for a given user, the n most similar users to a given user, the neighbourhood

of a given users with a specified similarity threshold, and finally a special

recommendation that provides a neighbourhood of most similar services to a given

service.

All recommendations are computed through the use of algorithms provided by

Taste libraries. Here a short list of algorithms we used (see the Taste documentation

for more details):

1. GenericItemBased: uses a given DataModel and ItemCorrelation to produce recommendations.

2. GenericUserBased: uses a given DataModel and UserNeighborhood to produce recommendations. (user similarity)

3. ItemAverage: always estimates preference for an Item to be the average of all known preference values for that Item. No information about Users is taken into account. (rank based)

4.IItemUserAverage: Like ItemAverageRecommender, except that estimated

preferences are adjusted for the Users' average preference value. For

example,   
say user X has not rated item Y. Item Y's average preference value

is 3.5. User X's average preference value is 4.2, and the average over all

preference values is 4.0. User X prefers items 0.2 higher on average, so, the

estimated preference for user X, item Y is 3.5 + 0.2 = 3.7.

5. TreeClustering: A Recommender that clusters Users, then determines the

clusters' top recommendations. This implementation builds clusters by

repeatedly merging clusters until only a certain number remain, meaning that

each cluster is sort of a tree of other clusters.

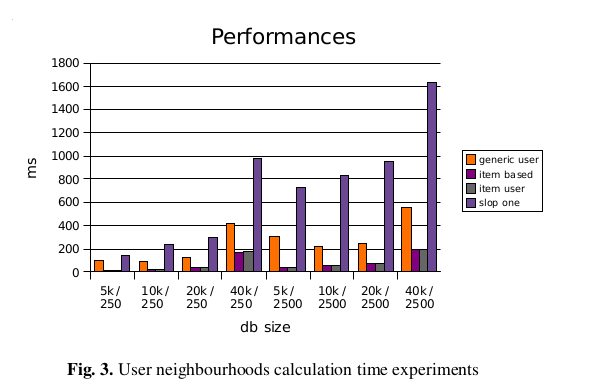
* Customer segmentation related

6. SlopeOne: A basic "slope one" recommender. This Recommender is

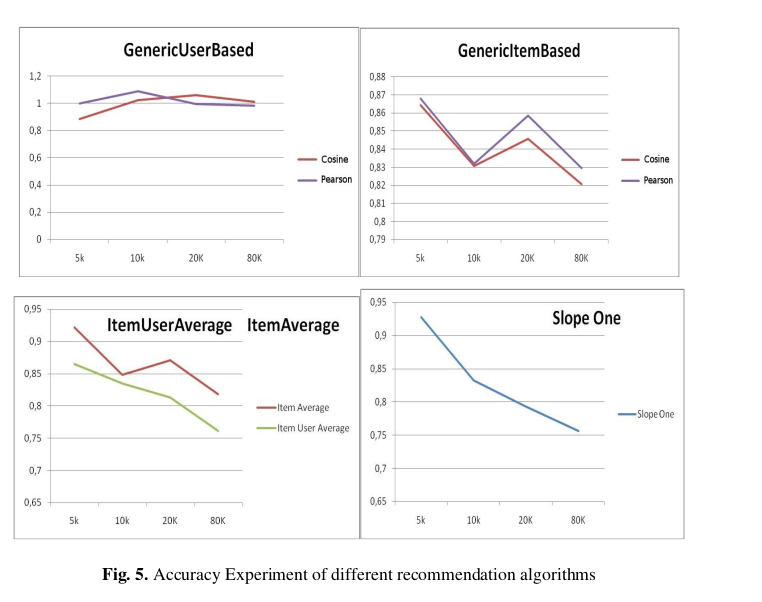
especially suitable when user preferences are updating frequently as it can

incorporate this information without expensive re-computation.

User segmentation based on interactions:



Accuracy of compared algos in above (1-6)



**Research Paper 04**

Use cases and challenges in telecom big data analytics

* **12 December 2016**
* **DOI:** [**https://doi.org/10.1017/ATSIP.2016.20**](https://doi.org/10.1017/ATSIP.2016.20)
* **chung-min chen**
* **Cite as:**

**Chen, C. (2016). Use cases and challenges in telecom big data analytics. APSIPA Transactions on Signal and Information Processing, 5, E19. doi:10.1017/ATSIP.2016.20**

**IMPORTANT - Challenges in Big data Analytics**

BA tools: Hadoop/ NoSQL/ Spark/ ML

A study reported that 75 of telecom operators surveyed would implement big data initiatives by 2017

(study: Heavy Reading: The Business Cases for Advanced Telecom Analytics,

Telecom Analytics World 2014, Atlanta, GA, Oct. 2014.)

Every operator is seeking new ways to increase operational efficiency and marketing effectiveness by leveraging big data technologies

Problems to tackle:

capabilities of uncovering insights from large volume of datasets?

What are the compelling use cases in telecom, and what are the challenges?

Earlier: Data warehousing/ On-line-analytical-processing (OLAP)/ data-mining

A suggested telcom framework:

framework contains three horizontal layers – resource, service, and customer, spanning across two vertical perspectives – infrastructure & product and operations

**\*\*Note:**

**Resource layer:** activities related to network build-out, planning, and monitoring. Operators constantly monitor performance of the networks (devices such as routers, switches, base stations, etc.) in order to assure smooth operation.

Data collected at this layer:

* alarms generated by the network devices
* key performance indicators (KPIs) such as
  + packet loss ratio,
  + latency,
  + traffic load, etc.

\* The datasets support :

* network planning,
* capacity management,
* fault management.

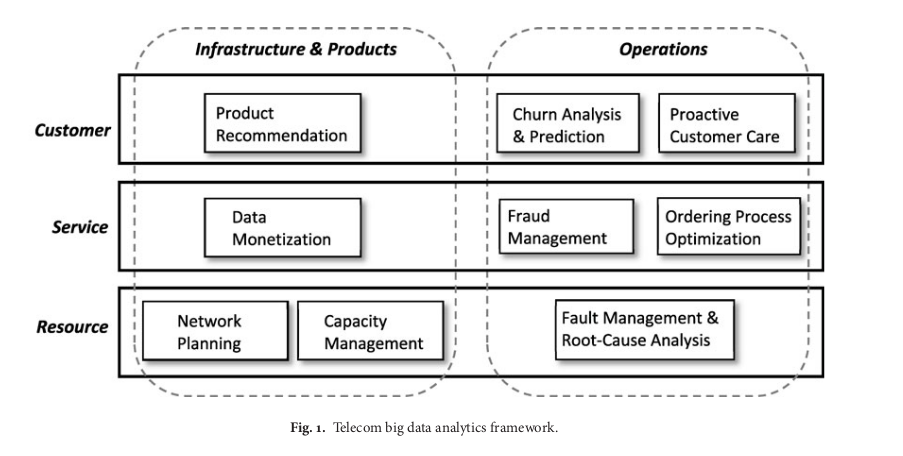
**service layer**: includes activities related to provisioning of user services (voice, data, and video). It also supports **proactive monitoring** and **reactive diagnostics** required by

service-level agreements – a contractual agreement between the operator and the users on the performance and availability of the subscribed services.

History logs from service provisioning can be used to improve the process,

Usage pattern data can be mined to detect frauds or monetized by selling to companies that are interested in reaching out to potential customers.

**customer layer:** the main task is Customer Relationship Management (CRM), which handles user inquires, orders, trouble tickets, and assure user satisfaction



**Churn analysis** predicts the probability that a user may terminate the service and provides insights on why the users are leaving.

Proactive customer care resolves issues the users may experience before they even know it by constantly monitoring the users’ quality of experience.

Since the mid-1990s, research in data warehousing, OLAP and data mining are abundant in the telecom application domain

A series of two workshops [8] focusing on database applications in telecom included research work on applying data warehousing and mining techniques to network resource and customer management.

[8] = Proceedings of VLDB International Workshop on Databases in

Telecommunications, I (Edinburgh, Scotland, UK, 1999, Lecture

Notes in Computer Science 1819, Springer 2000), II (Rome, Italy,

2001, Lecture Notes in Computer Science 2209, Springer 2001).

At the customer management layer, commercial solutions are available for churn analysis.

**IMPORTANT: Big Data Challenges:**

Variety:

Today, the proliferation of applications enabled by the Web, mobile networks, GPS, and social media has forever changed the horizon.

The numerous data points created and made avail- able by these applications have resulted in virtually a “data rainforest” with highly diverse sources of structured (tab- ular), semi-structured (objects, log records), unstructured (free text), and streaming data

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Informations Variety in Telcom Domain**

The information may include, e.g. what websites you visited, how much time you spent talking on the phone, watching video, and on OTT (over-the-top) apps such as Skye,

WhatsApp, and Facebook. By pairing these usage data with the network KPIs the operators are able to gain more insights into users’ QoE (Quality of Experience) on different services.

**\*\*\*\*\*\*\* Big data/ Capacity**

Telecom data has grown exponentially since the age of broadband and 3G. This trend

will continue as the networks evolve (optical, 4G/LTE, 5G), allowing users to access, contribute, and share ever increasing contents on the Internet.

**\*\*\*\*\* technology Availability**

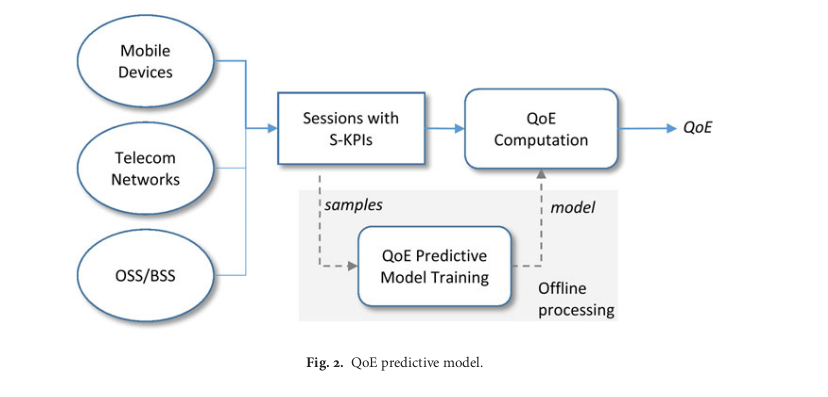
Many open sources have either a free community version or an enterprise version with moderate service/license fee. The Apache projects [13] are one of the most prominent examples.

**\*\*\*\*\*\* IMPORTANT \*\*\*\*\*\*\***

Needle in a haystack – how to uncover correlation and actionable insights from highly dimensional data space? The most challenging, and usually the most exciting, task

in any data analytics project is to identify the correlations among the variables (or features) and uncover the relationship between the variables and the metric to be predicted (labels)

Data integration and quality – your results are only as good as your data quality. Contrary to what most people think, data integration and cleansing usually take up the biggest chunk of time in any data analytics project. The tasks may involve traditional extract, transform, and load (ETL) and data reconciliation



In telecom, QoE is a measure of customer satisfaction on the service(s) she experienced. QoE can be service specific (e.g.,video QoE) or an overall measure across all services (e.g.,

video, voice and data altogether). QoE is usually measured on a scale of 1–10, though other objective metrics have also been proposed for service-specific QoE

Accurate measurement of customer QoE can help operators to predict customer churns and identify customers for service upgrade or target marketing.

**The data may include, for example, user locations, click stream logs, app usage, network performance, subscriber plans, and demography,and call center logs. Each user session (e.g., video, voice, or Web sessions) is associated with a set of service KPIs (S-KPIs) calculated based on the raw network data.**

Today Challenge: **How to predict subjective QoE (depend on user/ situation and interaction between user and relevant service) - ML**

**Machine Learning Techniques for Recommender systems**

**Research Paper 05 (Opinion mining/ context-aware rec systems)**

Using Opinion Mining in Context-Aware Recommender Systems: A Systematic Review

Camila Vaccari Sundermann 1, \* , Marcos Aurélio Domingues 2 , Roberta Akemi Sinoara 1 ,

Ricardo Marcondes Marcacini 3 and Solange Oliveira Rezende 1

28 January 2019

Context-aware recommender systems : Provide recommendations based on user-context

**There are increasing efforts to incorporate the rich information embedded in user’s reviews/texts into the recommender systems. Given the importance of this type of texts and their usage along with opinion mining and contextual information extraction techniques for recommender systems**

Traditional recommender systems focus on user and item data to generate recommendations.

Examples of such techniques include collaborative filtering, content-based filtering, and hybrid

approaches.

Collaborative filtering is a recommendation technique that finds correlations among users

or among items to generate recommendations

**However, empirical studies indicate that context-aware approaches can produce more precise recommendations**

Ref:

Adomavicius, G.; Tuzhilin, A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. Knowl. Data Eng. 2005, 17, 734–749. [CrossRef]

There are many definitions of context in the literature, depending on the application area [1]. In this work, the term context is defined as any information that can be used to characterize the situation of an entity (item or user)

One of the main challenges is the difficulty in the acquisition of contextual information. There is a lack of automatic methods for extracting this type of information.

Informations to extract from reviews/ comments:

Spects, which are the attributes of an item that a user discusses in a review; (ii) overall opinions, which can be represented by the orientations of users’ sentiments for items; (iii) aspect opinions, which are opinions about specific characteristics of an item; and (iv) contextual information

The interest of a user on an item is usually measured by a rating which can be obtained either

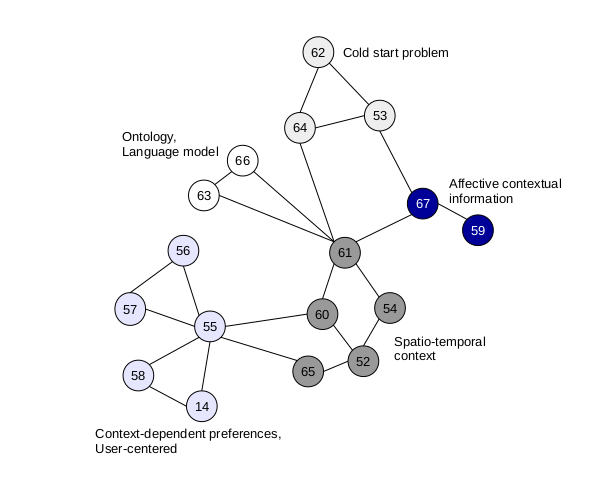
explicitly or implicitly.

* Numeric: when numerical values are assigned to products/services, for example, the five stars on the Amazon website
* Ordinal: when the user is prompted to select a term that best indicates his/her opinion on an item, such as “I agree”, “I am neutral” and “I disagree”
* Binary: when the user simply decides if an item is good or bad
* Unary: this kind of ratings was popularized by Facebook where users can mark his/her interest in a post or photo by clicking a button “Like”

Research Questions:

* What contextual information has been adopted for making recommendations?
* How has the contextual information been extracted?
* What opinion information has been adopted for making recommendations?
* How has the opinion information been extracted?
* Which textual sources have been used for the extraction of both context and opinion

Information?



No of researches in the area

**RESEARCH PAPER 06**

An enhanced ensemble classifier for telecom churn prediction

using cost based uplift modelling

* Ahmed, A.A.Q., Maheswari, D. An enhanced ensemble classifier for telecom churn prediction using cost based uplift modelling. *Int. j. inf. tecnol. 11*, 381–391. https://doi.org/10.1007/s41870-018-0248-3, 2019.

Regular classification approaches fail to effec-

tively predict churn due to low correlation levels between

conventional performance metrics and business goals

Example Intro:

In the last decade number of mobile phone users has

expanded exponentially. According to the United Nations

Department of Economic and Social affairs, the population

of the world which is estimated to be about 7.32 billion

people in 2016, would have more than 7? billion mobile

phone users with a penetration rate of 97% [1], nearly ten

times from the 738 million users recorded earlier in 2000.

Several western countries have already exhibited penetra-

tion rates above 100%, implying that the number of sub-

scriptions outnumber of the actual population.

**\*\*\*\***

acquiring new customers is five to six times costlier

than retaining existing customers. Furthermore, long-term

customers are much more profitable than scouting for new

customers. These old customers tend to introduce more

customers via referrals.

**Facts : Dialog**

Revenue

Rs. 120,142million (FY 2020)

Rs. 116,827million (FY 2019)

EBIDTA

Rs. 50,854million (FY 2020)

Rs. 46,703million (FY 2019)

Share in Issue Market Capitalisation

8,171,151,305

as at 31st December 2020 Rs. 101.3 billion

as at 31st December 2020

Customer base

16,287,445

Prepaid customers – 14,817,619 and postpaid customers – 1,469,826

(Status as at 31st December 2020)

Operational Network

Dialog operates over 5,700 (2G), 3,300 (3G) and over 3,800 (4G) base station sites distributed across all 9 provinces of Sri Lanka with a network coverage of approximately 90% of the country’s land mass and 98% of Sri Lanka’s population

**SLT: Facts**

Net Value: $326.5M

CEO: Lalith Mohan Seneviratne

The Sri Lanka Telecom Group provides diversified services and an entire range of ICT solutions that cover fixed and mobile telephony, broadband, data services, Internet Protocol Television (IPTV), cloud computing and hosting services, and networking solutions to its varied customers via the latest technologies. The SLT Group primarily focuses on three operating segments which are the Group’s strategic segments:

* Fixed ICT operations
* Mobile ICT operations
* Other segment operations

Vision:

All Sri Lankans seamlessly

connected with world-class

information, communication and

entertainment services.

Mission:

Your trusted and proven partner for

innovative and exciting communication

experiences delivered with passion,

quality and commitment

Fixed and mobile ICT operations

constitute the SLT Group’s core business.

In 2019, they collectively accounted for

98% of revenue, 98% of total assets, and

nearly 100% of capital expenditure of the

SLT Group.

Mobitel (Pvt) Ltd. (“Mobitel”), a

fully-owned subsidiary of SLT, offers

mobile ICT services including mobile

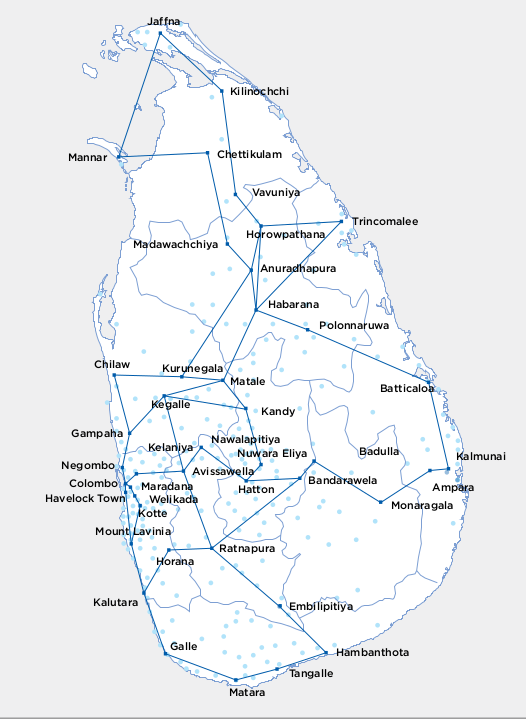
telephony services, high-speed

broadband, enterprise solutions,

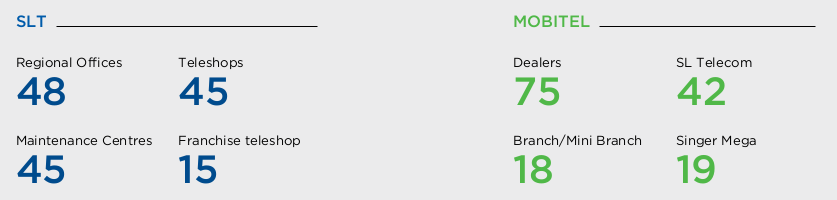
international services

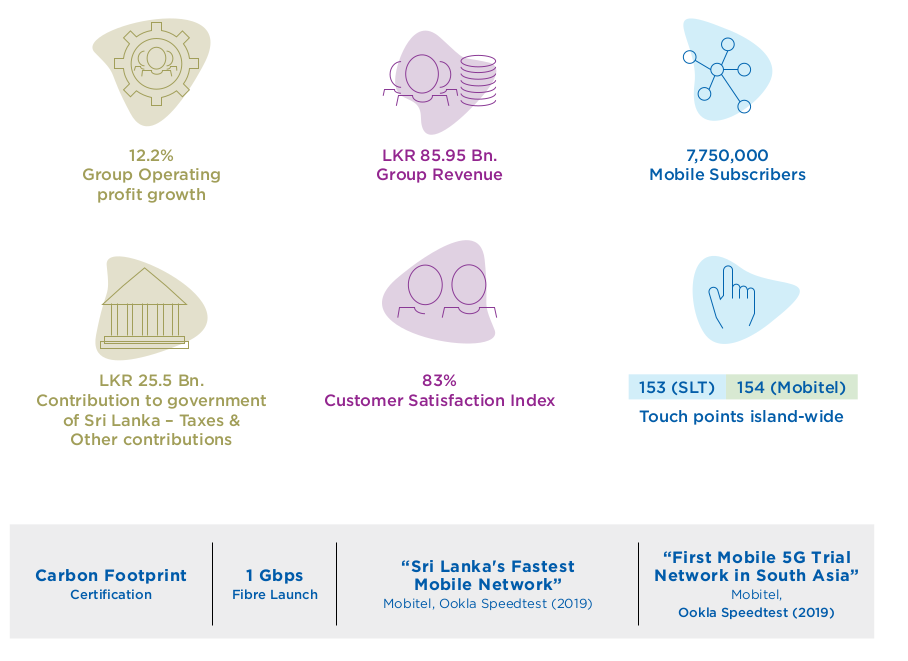
* Trial the first 5G network in South Asia
* international roaming with the partnership of a global web of over 650 networks
* Mobitel now connects all new smart devices with embedded SIMs, such as Apple iPhones and smartwatches, to facilitate the growth of new generation devices.
* Currently, 62% of Mobitel’s sites are connected with superfast fibre, making Mobitel the largest fibre connected mobile network in Sri Lanka

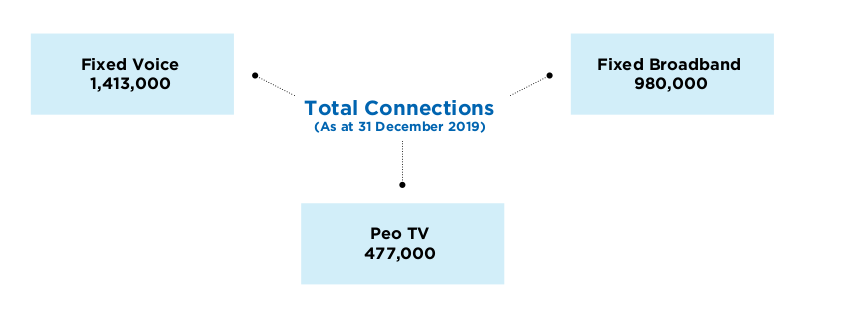
SLT coverage:



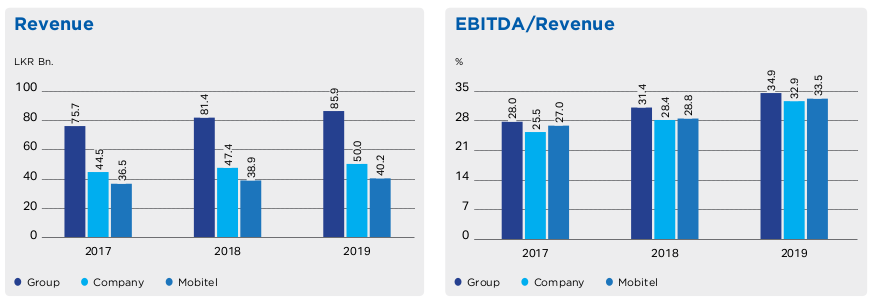
* South asia’s first submarine cable depot in Galle

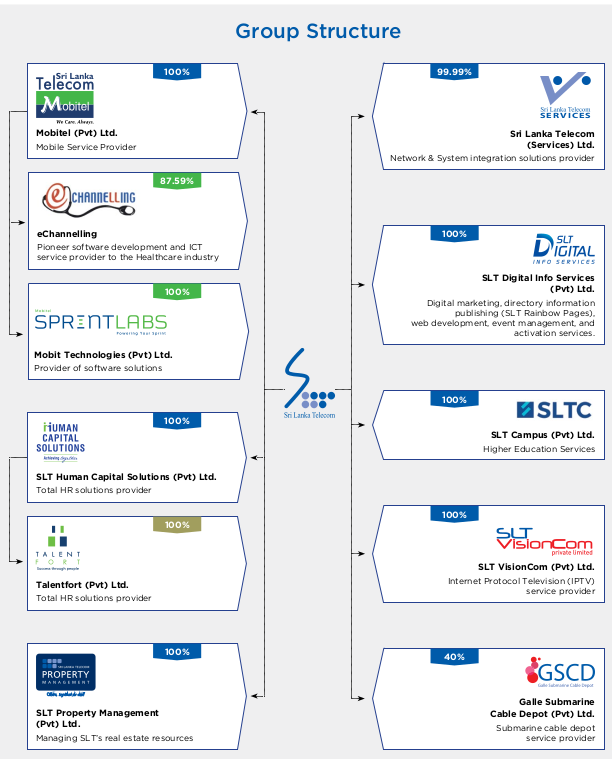






Revenue:





**Sri Lanka Stats (Digital Sri Lanka 2020)**

<https://datareportal.com/digital-in-sri-lanka>

## **Internet users in Sri Lanka**

* There were **10.10 million** internet users in Sri Lanka in January 2020.
* The number of internet users in Sri Lanka **increased** by **399 thousand** (+4.1%) between 2019 and 2020.
* Internet penetration in Sri Lanka stood at **47%** in January 2020.

## **Social media users in Sri Lanka**

* There were **6.40 million** social media users in Sri Lanka in January 2020.
* The number of social media users in Sri Lanka **increased** by **491 thousand** (+8.3%) between April 2019 and January 2020.
* Social media penetration in Sri Lanka stood at **30%** in January 2020.

## **Mobile connections in Sri Lanka**

* There were **31.80 million** mobile connections in Sri Lanka in January 2020.
* The number of mobile connections in Sri Lanka **increased** by **2.2 million** (+7.5%) between January 2019 and January 2020.
* The number of mobile connections in Sri Lanka in January 2020 was equivalent to **149%** of the total population.