Developing a Subscription-Likelihood Prediction Model for Rich Communication Services

Capstone Proposal

for

Udacity Machine Learning Engineer Nanodegree

By

Oluwasegun Francis Sogunle

February 17, 2019

1 Domain Background

Internet-based communications services now shape the way people communicate and collaborate. Consequently, Mobile Network Operators (MNOs) need to meet the ever-growing demands of their subscribers who use these services. However, the MNOs are not the sole providers of these internet-based services and more often than not, they do not have contracts with these third-party service providers such as Apple, Microsoft, Facebook and Google. In an attempt to retain their customers, the Global System for Mobile Communications Association (GSMA)[1] introduced the Rich Communications Services (RCS) [1] framework as a unified communications solution. RCS aims to provide services ranging from traditional voice call and Short Messaging Service (SMS) to instant messaging, voice and video over IP (VVoIP),[2] geolocation exchange, and presence, among others [3].

2 Problem Statement

Although RCS has been positioned to be at the core of MNO service offerings for future generation telecommunication networks, there has been relatively low adoption and pervasiveness of this suite of services [4]. Challenging issues such as the initial support from equipment manufacturers, creation of attractive billing models and well-tailored service offerings, and targeting the right and adequate set of customers contribute to this state. As a result, MNOs would benefit from services that could (as a first step), *predict the most likely set of customers* based on their Call Detail Records (CDRs).[3]

The likely set of customers in this context are those who are active users of the telecommunication network services at a specific location. In other words, those who extensively use the current voice call and SMS services that MNOs offer. As a rule, these class of customers are likely to adopt the new RCS service offerings. Consequently, MNOs could leverage this information to make further decisions. Take, for instance, creating location-specific or user-specific profiles for RCS service offerings. That is, without techniques that

[1] A consortium of MNOs, Original Equipment Manufacturers (OEMs), and telecommunication-inclined stakeholders

[2] IP - Internet Protocol

[3] CDR is a log of a customer's activities while using an MNO's communication infrastructure. For example, the network generates a CDR when a customer places a voice call.

reliably identify these active customers, making such higher-level decisions may be difficult and can prove ineffective since there is no focus on the class of subscribers that are likely to adopt RCS.

3 Datasets and Inputs

The proposed project will adopt mobile phone activity [2] dataset available on Kaggle. Considering the length of records in each file at the repository, this project will utilise records found in **sms-call-internet-mi-2013-11-01.csv** that was gathered within a day. The dataset is an MNO-generated CDR for subscribers in Milan. It will serve as a sample of the large dataset spanning a week. The file will be split for training, testing and validation. The file's Metadata which will be used to build the predictive model is as follows:

- **CellID**: identification string of a given square of Milan grid.
- **countrycode**: A nation's country code for the subscriber's number.
- **smsin**: Amount of SMSs received a given square id and during a given time interval.
- **smsout**: Amount of SMSs sent inside a given square id during a given time interval.
- **callin**: Amount of calls received inside the square id during a given time interval.
- **callout**: Amount of issued calls inside a given square id during a given time interval.
- **internet**: Number of CDRs generated inside a given Square id during a given time interval.

Notably, the square grid identifies the location of a specific cell which is identified by a CellID. The dataset comprises 1891928 records and 8 features. However, the **datetime** feature will not be used as it retains a constant value across records.

4 Solution Statement

The proposed solution will adopt an hybridization of supervised and unsupervised learning algorithms. To begin, the dataset is without labels for the considered *two classes* of data. That is, active subscribers and other subscribers, which could be partial or non-active users. In other to make this distinction, the Gaussian Mixture Model (GMM) clustering algorithm will be used to label the data. Thereafter, a well-trained artificial neural network[4] would be used to create a model checkpoint that makes predictions on unseen customer CDRs in order to determine their likelihood of subscribing to RCS services. In other words, the output of the model would be a likelihood (i.e., probability) of subscription to RCS services.

5 Benchmark Model

Taking into consideration that this is a novel proposition for RCS service providers in the telecommunications space, the benchmark model could be one without machine learning. However, since the labels for the data had to be derived with a clustering technique, a possible benchmark model would be to use a supervised learning algorithm such as logistic regression to fit the data. Thereafter, the proposed hybridized model would be compared with the GMM-based regression model, which will serve as the benchmark. In essence, examining the performance of the neural network and logistic regression algorithms on the GMM-clustered data.

6 Evaluation Metrics

The proposed model and its benchmark will be examined with the F-beta score and Receiver Operating Characteristic (ROC) curve. These metrics will present the basis for the model evaluation and provide likely pointers and insights for possible improvements on the project. In essence, the models would be examined and compared on their ability to reliably identify the target RCS customers when given an instance of a CDR.

[4]Either a Convolutional Neural Network (CNN) or a multi-layer perceptron

7 Project Design

The workflow for the proposed project is outlined as follows:

- Explore and pre-process data
  - Understand the dataset structure.
  - Remove records with NaN values.
  - Scale dataset features.
  - Remove outliers.
  - Examine feature relevance and reduce dimensions through Principal Component Analysis (PCA).
- Define the ground truth for the data set through clustering
  - Investigate optimal value for number of clusters using the silhouette score.
  - Extract labels from the clustering result.
  - Validate and analyse the clustering result.
- Implement the proposed hybrid model.
  - Build an artificial neural network on the clustering result.
- Evaluate models
  - Implement logistic regression on the clustering result as a benchmark model.
  - Compare the F-beta score and ROC curves for both models (i.e., proposed and benchmark).

References

[1] GSMA. *The RCS Ecosystem - Future Networks*. Available at: `https://www.gsma.com/futurenetworks/rcs/the-rcs-ecosystem/`. [Accessed 16 February, 2019].

[2] Kaggle. *Mobile phone activity in a city*. Available at: `https://www.kaggle.com/marcodena/mobile-phone-activity`. [Accessed 16 February, 2019].

[3] *RCS Universal Profile Service Definition Document*. RCC.71. Technical Specification. Version 2.2. GSMA. May 2018.

[4] Sogunle, O. F. *A Unified Data Repository for Rich Communication Services*. Master's Thesis. Rhodes University, South Africa, April 2017.