Abstract

Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenues of the companies, especially in the telecom field, companies are seeking to develop means to predict potential customer to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn. The main contribution of our work is to develop a churn prediction model which assists telecom operators to predict customers who are most likely subject to churn. The model developed in this work uses machine learning techniques on big data platform and builds a new way of features' engineering and selection. In order to measure the performance of the model, the Area Under Curve (AUC) standard measure is adopted, and the AUC value obtained is 93.3%. Another main contribution is to use customer social network in the prediction model by extracting Social Network Analysis (SNA) features. The use of SNA enhanced the performance of the model from 84 to 93.3% against AUC standard. The model was prepared and tested through Spark environment by working on a large dataset created by transforming big raw data provided by SyriaTel telecom company. The dataset contained all customers' information over 9 months, and was used to train, test, and evaluate the system at SyriaTel. The model experimented four algorithms: Decision Tree, Random Forest, Gradient Boosted Machine Tree "GBM" and Extreme Gradient Boosting "XGBOOST". However, the best results were obtained by applying XGBOOST algorithm. This algorithm was used for classification in this churn predictive model.

Keywords: Customer churn prediction, Churn in telecom, Machine learning, Feature selection, Classification, Mobile Social Network Analysis, Big data

Introduction

The telecommunications sector has become one of the main industries in developed countries. The technical progress and the increasing number of operators raised the level of competition [1]. Companies are working hard to survive in this competitive market depending on multiple strategies. Three main strategies have been proposed to generate more revenues [2]: (1) acquire new customers, (2) upsell the existing customers, and (3) increase the retention period of customers. However, comparing these strategies taking the value of return on investment (RoI) of each into account has shown that the third strategy is the most profitable strategy [2], proves that retaining an existing customer costs much lower than acquiring a new one [3], in addition to being considered much easier than the upselling strategy [4]. To apply the

third strategy, companies have to decrease the potential of customer's churn, known as "the customer movement from one provider to another" [5].

Customers' churn is a considerable concern in service sectors with high competitive services. On the other hand, predicting the customers who are likely to leave the company will represent potentially large additional revenue source if it is done in the early phase [3].

Many research confirmed that machine learning technology is highly efficient to predict this situation. This technique is applied through learning from previous data [6, 7].

The data used in this research contains all customers' information throughout nine months before baseline. The volume of this dataset is about 70 Terabyte on HDFS "Hadoop Distributed File System", and has different data formats which are structured, semi-structured, and unstructured. The data also comes very fast and needs a suitable big data platform to handle it. The dataset is aggregated to extract features for each customer.

We built the social network of all the customers and calculated features like degree centrality measures, similarity values, and customer's network connectivity for each customer. SNA features made good enhancement in AUC results and that is due to the contribution of these features in giving more different information about the customers.

We focused on evaluating and analyzing the performance of a set of tree-based machine learning methods and algorithms for predicting churn in telecommunications companies. We have experimented a number of algorithms such as Decision Tree, Random Forest, Gradient Boost Machine Tree and XGBoost tree to build the predictive model of customer Churn after developing our data preparation, feature engineering, and feature selection methods.

There are two telecom companies in Syria which are SyriaTel and MTN. SyriaTel company was interested in this field of study because acquiring a new customer costs six times higher than the cost of retaining the customer likely to churn. The dataset provided by SyriaTel had many challenges, one of them was unbalance challenge, where the churn customers' class was very small compared to the active customers' class. We experimented three scenarios to deal with the unbalance problem which are oversampling, undersampling and without re-balancing. The evaluation was performed using the Area under receiver operating characteristic curve "AUC" because it is generic and used in case of unbalanced datasets [8].

Many previous attempts using the Data Warehouse system to decrease the churn rate in SyriaTel were applied. The Data Warehouse aggregated some kind of telecom data like billing data, Calls/SMS/Internet, and complaints. Data Mining techniques were applied on top of the Data Warehouse system, but the model failed to give high results using this data. In contrast, the data sources that are huge in size were ignored due to the complexity in dealing with them. The Data Warehouse was not able to acquire, store, and process that huge amount of data at the same time. In addition, the data sources were from different types, and gathering them in Data Warehouse was a very hard process so that adding new features for Data Mining algorithms required a long time, high processing power, and more storage capacity. On the other hand, all these difficult processes in Data Warehouse are done easily using distributed processing provided by big data platform.

Fig. 1 Hortonworks data platform HDP—big data framework

of HDP, where each group of tools is categorized under specific specialization like Data Management, Data Access, Security, Operations and Governance Integration.

The installation of HDP framework was customized in order to have the only needed tools and systems that are enough to go through all phases of this work. This customized package of installed systems and tools is called SYTL-BD framework (SyriaTel's big data framework). We installed Hadoop Distributed File System HDFS[2] to store the data, Spark execution engine[3] to process the data, Yarn[4] to manage the resources, Zeppelin[5] as the development user interface, Ambari[6] to monitor the system, Ranger[7] to secure the system and (Flume[8] System and Scoop[9] tool) to acquire the data from outside SYTL-BD framework into HDFS.

The used hardware resources contained 12 nodes with 32 Gigabyte RAM, 10 Terabyte storage capacity, and 16 cores processor for each node. A nine consecutive months dataset was collected. This dataset will be used to extract the features of churn predictive model. The data life cycle went through several stages as shown in Fig. 2

Spark engine was used in most of the phases of the model like data processing, feature engineering, training and testing the model since it performs the processing on RAM. In addition, there are many other advantages. One of these advantages is that this engine containing a variety of libraries for implementing all stages of machine learning lifecycle.

Data acquisition and storing

Moving the data from outside SYTL-BD into HDFS was the first step of work. The data is divided into three main types which are structured, semi-structured and unstructured.

Apache Flume is a distributed system used to collect and move the unstructured (CSV and text) and semi-structured (JSON and XML) data files to HDFS. Figure 3 shows

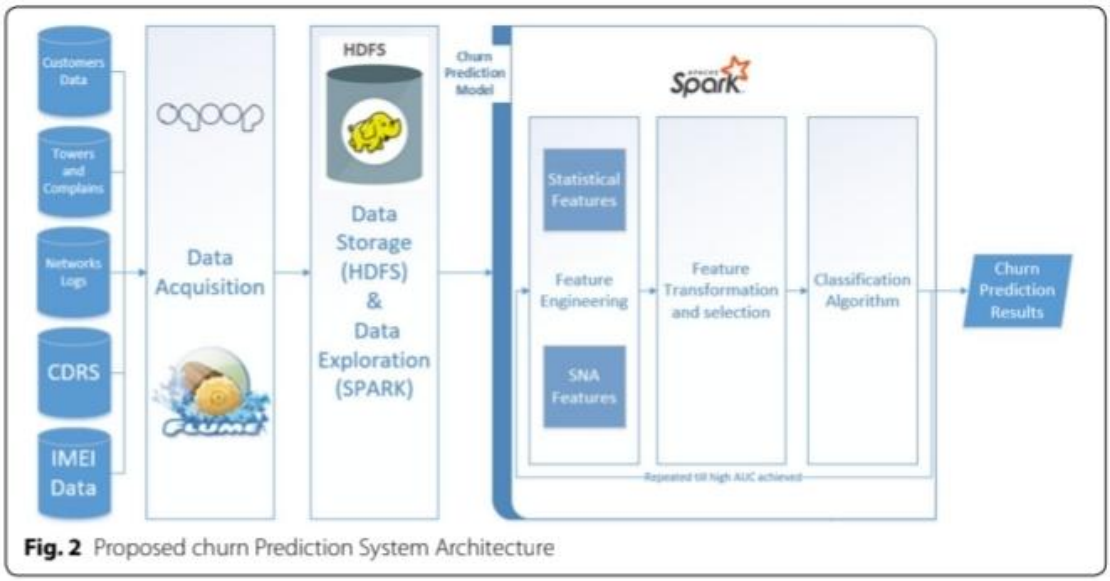

Fig. 2 Proposed churn Prediction System Architecture

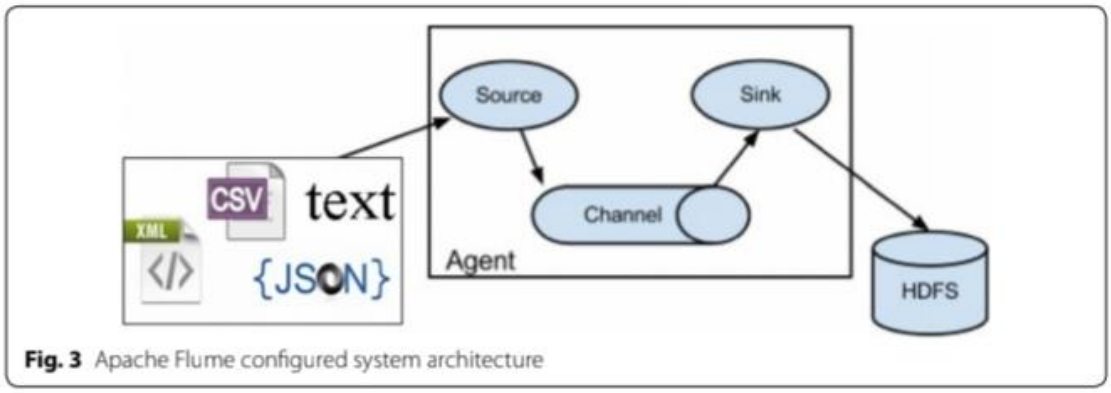

Fig. 3 Apache Flume configured system architecture

the designed architecture of flume in SYTL-BD. There are three main components in FLUME. These components are the data Source, the Channel where the data moves and the Sink where the data is transported.

Flume agents transporting files exist in the defined Spooling Directory Source using one channel, as configured in SYTL-BD. This channel is defined as Memory Channel because it performed better than the other channels in FLUME. The data moves across the channel to be finally written in the sink which is HDFS. The data transformed to HDFS keep in the same format type as it was.

Apache SQOOP is the distributed tool used to transfer the bulk of data between HDFS and relational databases (Structured data). This tool was used to transfer all the data which exists in databases into HDFS by using Map jobs. Figure 4 shows the architecture of SQOOP import process where four mappers are defined by default. Each Map job selects part of the data and moves it to HDFS. The data is saved in CSV file type after being transported by SQOOP to HDFS.

After transporting all the data from its sources into HDFS, it was important to choose the appropriate file type that gives the best performance in regards to space utilization and execution time. This experiment was done using spark engine where

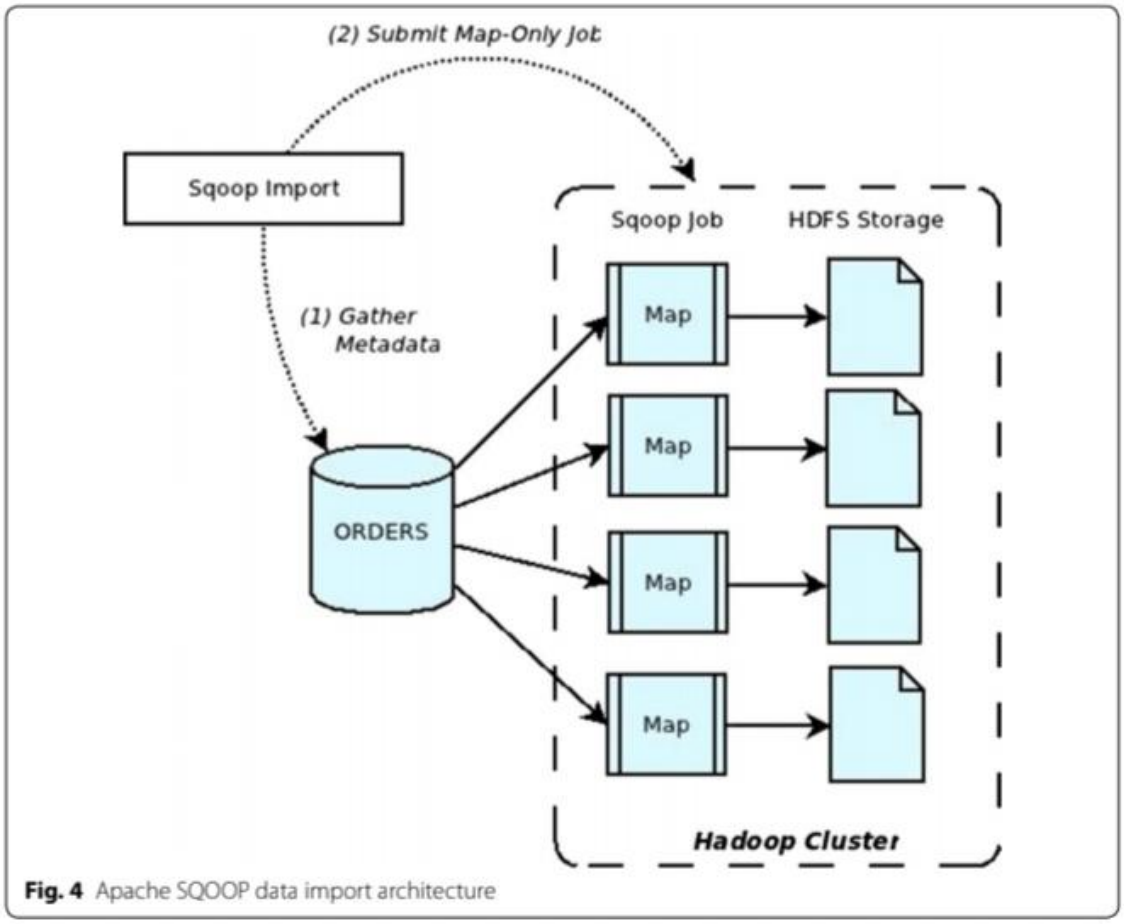

Fig. 4 Apache SQOOP data import architecture

Data Frame library[10] was used to transform 1 terra byte of CSV data into Apache Parquet[11] file type and Apache Avro[12] file type. In addition to that, three compression scenarios were taken into consideration in this experiment.

Parquet file type was the chosen format type that gave the best results. It is a columnar storage format since it has efficient performance compared with the others, especially in dealing with feature engineering and data exploration tasks. On the other hand, using Parquet file type with Snappy Compression technique gave the best space utilization. Figure 5 shows some comparison between file types.

Feature engineering

The data was processed to convert it from its raw status into features to be used in machine learning algorithms. This process took the longest time due to the huge numbers of columns. The first idea was to aggregate values of columns per month (average, count, sum, max, min ...) for each numerical column per customer, and the count of distinct values for categorical columns.

Another type of features was calculated based on the social activities of the customers through SMS and calls. Spark engine is used for both statistical and social features, the library used for SNA features is the Graph Frame.

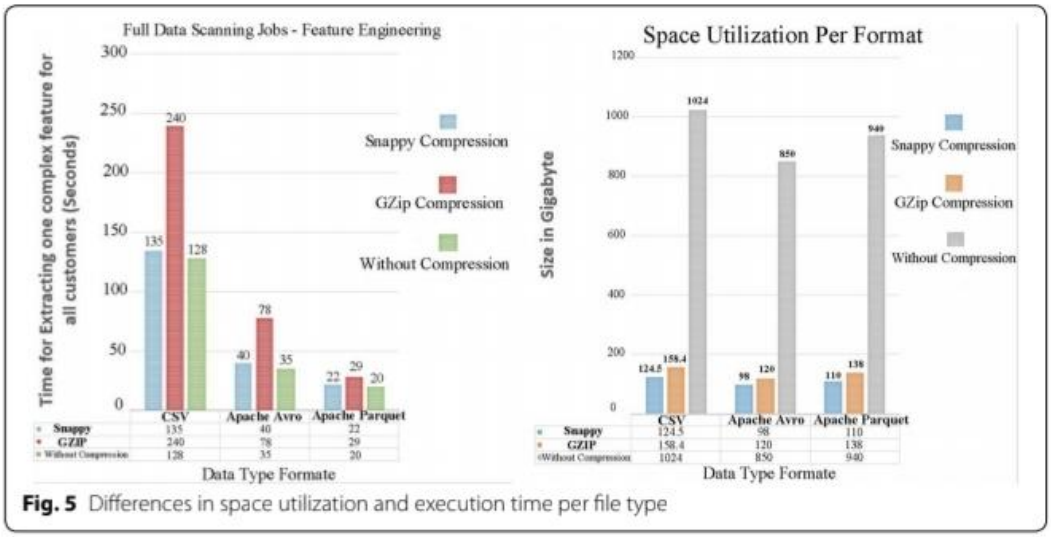

Fig. 5 Differences in space utilization and execution time per file type

- *Statistics features* These features are generated from all types of CDRs, such as the average of calls made by the customer per month, the average of upload/download internet access, the number of subscribed packages, the percentage of Radio Access Type per site in month, the ratio of calls count on SMS count and many features generated from aggregating data of the CDRs.

  Since we have data related to all customers' actions in the network, we aggregated the data related to Calls, SMS, MMS, and internet usage for each customer per day, week, and month for each action during the nine months. Therefore, the number of generated features increased more than three times the number of the columns. In addition, we entered the features related to complaints submitted from the customers from all systems. Some features were related to the number of complaints, the percentage of coverage complaints to the whole complaints submitted, the average duration between each two complaints sequentially, the duration in "Hours" to close the complaint, the closure result, and other features.

  The features related to IMEI data such as the type of device, the brand, dual or mono device, and how many devices the customer changed were extracted.

  We did many rounds of brainstorming with seniors in the marketing section to decide what features to create in addition to those mentioned in some researches. We created many features like percentage of incoming/out-coming calls, SMS, MMS to the competitors and landlines, binary features to show if customers were subscribing some services or not, rate of internet usage between 2G, 3G and 4G, number of devices used each month, number of days being out of coverage, percentage of friends related to competitor, and hundred of other features.

  Figures 6 and 7 visualize some of the basic categorical and numerical features to give more insight on the deference between churn and non-churn classes.

2 RELATED WORK

Research in the area of customer churn is always a trending topic. The unbridled growth of databases in recent years brings data mining to the forefront of new business technologies., becoming our only hope for elucidating the patterns that underlie it [23]. Significant research in the field of churn prediction is being carried out using various statistical and data mining techniques since a decade. This chapter presents the recent and prominent publications on churn prediction in the recent years.

Gavril Toderean et al [8] presented an advanced data mining methodology that predicts customer churn in the pre-paid mobile telecommunications industry using call detail records dataset that consists of 3333 customers with 21 attributes each and a churn dependent variable with two classes Yes/No. Few attributes include the information about their corresponding inbound/outbound SMS count and voice mail. A principal component algorithm was applied to reduce the dimensionality of data and to eliminate the problem of multicollinearity. Three machine learning algorithms, namely neural networks, support vector machines and Bayesian networks were used to predict churn variable based independent variables. These models were evaluated using confusion matrix, gain measure and ROC curve. An overall accuracy of 99.10%, 99.55% and 99.70% were achieved for Bayesian networks, neural networks and support vector machines respectively.

Kiran Dahiya et al [9] proposed a new framework for churn prediction model, implemented it using WEKA data mining software. Each customer was classified as a potential churner or non-churner. The framework discussed was based on Knowledge Discovery Data process. Three different datasets, small, medium and large with varying attributes were considered. The efficiency and performance of decision tree and logistic regression techniques have been compared. Accuracy achieved with decision tree was much greater than logistic regression.

Utku Yabas et al [10] explains about subscriber churn analysis and prediction for mobile and wireless service providers. A real and complied dataset by Orange Telecom, 2009 was used. Main emphasis was laid on ensemble methods that encompass single methods to improve the solution to churn prediction problem. These results were compared with that of meta-classifiers, namely logistic regression, decision trees and random forests; and had encouraging values when considered for both ROC score and computing efficiency.

Saad Ahmed Qureshi et al [1] aims to present commonly used data mining techniques for churn prediction. The dataset used was obtained from Customer DNA website and contains traffic data of 1,06,000 customers and their usage behavior for three months. The class imbalance problem was solved by re-sampling. Regression analysis, Artificial Neural Networks, K-Means Clustering, Decision Trees including CHAID, Exhaustive CHAID, CART and QUEST were taken into consideration to identify churn. The results were compared based on the values of precision, recall and F-measure. Decision trees, especially Exhaustive CHAID were found to be the most accurate algorithm in identifying potential churners.

Muhammad Raza Khan et al [5], presented a unified analytic framework for detecting the early warnings of churn, and assigning a "Churn Score" to each customer that indicates the likelihood of a particular customer to churn within a predefined amount of time. The approach uses a brute force approach to feature engineering that generates a large number of overlapping features from customer transaction logs, then uses two related techniques to identify the features and metrics that are most predictive of customer churn. These features are then fed into a series of supervised learning algorithms that can accurately predict subscriber churn. For a dataset of roughly 1,00,000 subscribers from a South Asian mobile operator observed for 6 months, an approximate of 90 percent accuracy was achieved.

In order to solve the problem of big customer churn of about 5.23 million customers from China Telecom and China Netcom for fixed communication network operators, Yue He et al [11], proposed a prediction model based on RBF neural network. It then subdivides the customers by Analog Complexion Cluster to guide and help manage marketing and related work.

Genetic Programming (GP) based approach along with AdaBoost for modeling the challenging churn problem was proposed by Adnan Idris et al [7]. The GP's evolution process was exploited by integrating an AdaBoost style boosting to evolve multiple programs per class and final predictions are made on the basis of weighted sum of outputs of GP programs. This was tested on two standard datasets, one by Orange Telecom and the other by cell2cell. The accuracy achieved was 89% for cell2cell dataset and 63% for the other.

Xiaohang Zhang et al [12], investigated the effects of network attributes on the accuracy of churn prediction. Network attributes refer to the interaction among customers and the topologies of their social network, which is constructed by the customer calling behaviors. The predictions of traditional attribute-based models, network attribute-based models and combined attributes models are compared and found that incorporating network attributes into predicting models can greatly improve the prediction accuracy. The network attributes can be useful complements to the traditional attributes.

Michael J.Prez et al [13] proposed to identify customers with service failures and determine the propensity for a customer to disconnect based on the frequency of a recent service failure reported and success of repair. The dataset used in this study was from monthly statistical reports of a national multi-system operator in the telecommunications industry over a 10month period during January to October 2008. Two approaches were used in this study. The first looked at the service experience of customers with a service failure, from provider's "phone survey statistics" of current customers with a service failure. The second approach looked at the frequency of customers who had disconnected their services following a service failure within a 30 day (monthly) reporting period, using empirical data from the telecommunications provider's "billing system". The proceedings stated that the customers subscribed for the triple-play of voice, video and internet access were more likely to cancel all services after a service failure than other customers.

In paper [14] by L.Bin et al, call details of 6000 customers of Personal Handy phone System Service in China are observed for 180 days. After data pretreatment, data of 4799 customers was preserved. In order to build an effective and accurate model, three experimentations were considered to improve the ability of churn prediction. These include: changing sub-periods for training data sets, changing misclassification cost in churn model, changing sample method for training data sets. The results suggested that these churn models have excellent performance, quite effective and feasible only for limited information and skewed class distribution.

In paper [6] by A.Idris and A.Khan, a dataset of 40,000 instances provided by cell2cell Telecom Company was pre-processed to a balanced form. In the preprocessing stage, in order to provide discriminating features to the classifiers mRMR, Fisher's ratio and F-Score feature extraction methods were used. For each of these methods, a linear search is performed to select the features which provide maximum discriminating information to the classifiers and hence produce better performance. When a linear search is performed for all the methods with rotation forest, for mRMR the accuracy for predicting the churners was 76.2%, while it was 69.1% and 65.2% for Fisher's ratio and F-Score respectively. For Random Forest, the accuracy of churn prediction for mRMR, Fisher's Ratio and F-Score were 74.2%, 71.6% and 71.3% respectively.