EasyChair Preprint

№ 683

Practical Implementation of Machine Learning and Predictive Analytics in Cellular Network Transactions in Real Time

Dahj Muwawa Jean Nestor and Kingsley A. Ogudo

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

December 17, 2018

Practical Implementation of Machine Learning and Predictive Analytics in Cellular Network Transactions in Real Time

***Abstract*—In order to keep a high revenue stream, Communication Service Providers in general, Network Mobile Operators specifically need to ensure a good level of customer satisfaction by assigning a big weight on the user's Quality of Experience (QoE). With billions of transactions done by customers on both voice and data daily, Communication Service Providers (CSPs) shift the focus in studying customer behavior and data patterns to pinpoint opportunities to improve customer services, service quality and predict when customers are likely to terminate contracts, to perhaps move to another CSP. CSPs have managed to build efficient IT infrastructures to store customer transactions. These exist in many forms such as file systems, databases, etc. In this paper, a simplified predictive analytics is done using the (Customer Relationship Management) CRM information records to classify potential customers likely to terminate their contracts, using logistic regression and random forest models. The paper describes the process to build a simple predictive models to apply on a telecoms dataset.**

Keywords—Machine Learning, Predictive Analytics, CRM, Logistic Regression, Random Forest, CSP, Telecommunications, Artificial Intelligence (AI).

I. INTRODUCTION

The rise of Machine Learning and Artificial intelligence is becoming significant in the area of telecommunications; from simple descriptive analysis, to more complex predictive models, CSPs are determined to find solutions to reduce customers' contract termination, referred to as churn, to find patterns of customer behavior which can help in improving retention and satisfaction. In the area of Sales & Marketing, CSPs' objective is to meet specific commercial goals and do a customer profile comparison to increase revenue by individualizing sales packages according to customers. The field of data mining and predictive analytics is situated in the intersection of Computer Science, Statistics and Machine Learning. Mining Telecoms data is about finding useful information in the Millions of customer transactions and contract information stored in large databases such as the CRM, in a structured and unstructured way [1]. The aim of this paper is to uncover processes that are required to do predictive analytics on CRM data and classify customers that are likely to churn in the near future. The Customer Relation Management (CRM) can be looked at as a big warehouse of structured customers' information. Based on existing well known Machine Learning algorithms Regression Models and Random Forest, CRM information is pre-processed and fed to a model to predict churn.

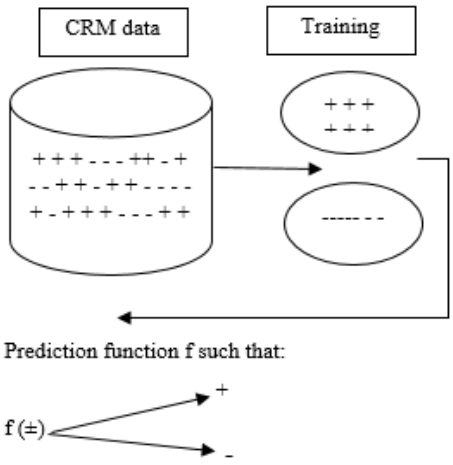

Figure 1: Predictive Analytics' Creed

Data collected from the CRM platform is used to predict customers likely to churn. The data, based on probability is divided to obtain the training set which is contains different characteristics of the data from which the model learn the behavior. A new set of CRM data is then fed to the prediction function to predict if a customer will churn.

In order to build a good predictive model, components of the predictors need to be evaluated and understood. Hence, a need of knowledge on both Machine Learning and Telecommunications. For this research, R platform is used [2] as it constitutes a solid platform for data analysis and statistics and is an open source platform.

II. PROBLEMATIC

A. The Prediction Effect

Running predictive analytics requires knowledge and expertise in both Machine Learning area and the domain in which the Analytics is being applied. Machine Learning is no magic in the sense that human intervention is still required to make the models efficient and reliable, including parameter tuning and evaluation of models. Predicting is not easy especially when dealing with future predictions. A typical illustration is that weather prediction is accurate at about 50% [3]. Comparing to statistical probability and guessing, Predicting brings real business value even in its lowest accuracy [4].

Predictive Analytics in the area of this research paper, aims to predict future behavior of customers in the

Telecommunications industry in order for the CSP to make business decisions and optimize the network.

B. Customer Relationship Management Limitation

The CRM contains contractual data for customers and aggregated statistical usage of network services for every single subscribers. Although this seems to be a rich customer data warehouse, it does not give a good insight on Quality of Experience (QoE) of each subscriber. To optimize on the QoE, more data management systems are required including probes, Performance Management, Fault management, and mostly CDRs (Call Data Records).

As mentioned in the above section, it is still necessary to run Predictive Analytics on CRM data to study the customer behavior and build an optimized marketing strategy.

III. CONTRIBUTION OF THE STUDY

On the topic of Machine Learning and Predictive Analytics is associated Big Data. The rise on the amount of data generated by people and machine is the modern day's main driving force for Big Data Analytics and Machine Learning. In data prediction, the size of the data needs to be taken into account. Running a prediction on 100 MB dataset is different from using 10TB of data. More computational resources are required to deal with Big Data in its entire form [5]. The contribution of the paper is to enlighten CSPs to create efficient, cost-effective Predictive Analytics environments to enhance business decisions in a cost effective manner and save on expensive Hardware and tools. There are currently several prediction algorithms and models which are being built for research and practical purposes, which leaves the area still an open area for improvement and adaptation to suit different business models and industries.

Big Data and Machine Learning are transforming health care and medical practices, leading advertisement and driving business decision [6]. Making Big Data and Machine Learning the driving force of Telecommunications would change the way Customer Experience is handled and will give competitive advantage.

IV. METHODOLOGY & DESIGN

Tackling a Machine Learning and Predictive Analytics problem depends on certain number processes and methodologies. The first step is to define the question of the Analytics. What is being predicted and what is used to predict it? The second step is the collection of data to be used for prediction. The third step is to understand and explore the data to build the necessary prediction features; the forth step is the choice of the algorithms to use for prediction; the fifth step is the parameter tuning and training of the datasets to discover patterns of customer behavior; the last step is to evaluate the model by testing it with a new dataset. The methodology and study design is illustrated in the Figure 2.

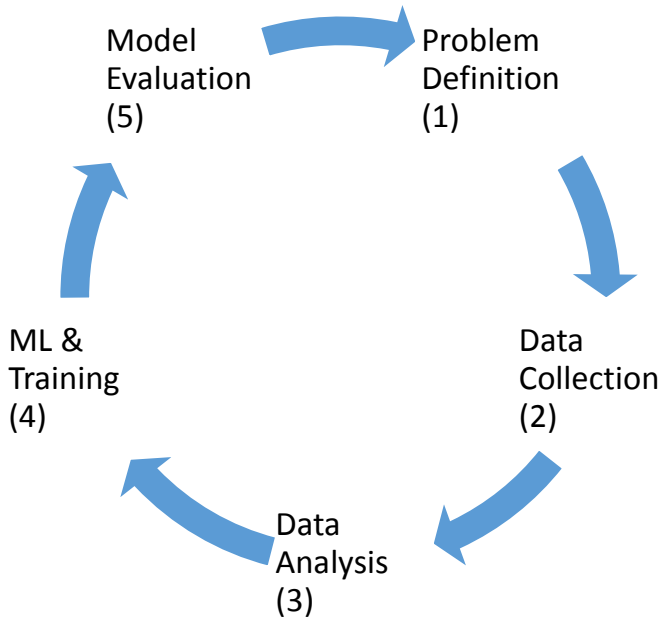

Figure 2 Predictive Study Design & Methodology

V. PROBLEM DEFINITION AND OBJECTIVES

Is it possible to classify or point out subscribers that are likely to terminate their contract with a certain CSP? Such is the general question to define the scope of the Analytics. Based on quantitative and qualitative data from CRM, we want to be able to predict if a certain customer will terminate his contract or not. The objectives of the Analytics are:

- Do an Exploratory Data Analysis on CRM information to examine relationships between different variables or predictors.
- Predict the outcome of subscribers, using the Algorithm with the best accuracy.
- Find based on decision tree, the weight of predictors that have decided on the outcome of the model.

VI. DATA COLLECTION

Information is stored in the CRM database in a structured way and collected in a systematic way either through API (Access Programming Interface) or through File management FTP (File Transfer Protocol) client. It is not possible to presume that CSP's data are stored in a single location for easy access, they are stored in different IT infrastructure depending on the transactions to be stored. The raw data used in this research paper are captured through FTP, coming from an Operator's CRM system.

A. Data Pre-processing & Cleaning

The raw data collected from the CRM has not been processed yet and might contain Null values and parameters that can affect

the prediction models: these are outliers, missing values and incorrect fields. The objective in this phase is to reduce the amount of Garbage in the models and understanding the predictors. The variables are as follow after preprocessing:

- SERVICE_NUMBER (Char): The Mobile Station International Subscriber Directory Number. This is the Mobile number.
- SERVICE_LINE (Char): the types of contract the subscriber is under.
- SERVICE_PLAN (Char): the customer service plan.
- SUBSCRIBER_NUMBER (Char): Unique ID to identify the Subscriber in the Network.
- SUBSCRIBER_STATUS (Char): the current state of the customer in the network.
- ACTIVATION_MONTH (Char): the activation date and month of the subscriber's contract.
- REGION (Char): the region where the subscriber is based. The region of the contract initiated.
- GENDER (Char): Indicates the gender of the subscribers.
- CUSTOMER_ID (Char): Unique number to identify the subscriber in the CRM system.
- INT_CALLS (INT): the number of International calls made and received.
- INT_CALLS_REVENUE (Decimal): Revenue generated on the International transactions.
- CALLS_TOTAL_NUMBER (INT): the total number of calls including both national and international.
- CALL_REVENUE (Decimal): Revenue generated on all the calls.
- CHARGEABLE_DURATION (Decimal): The amount of time billed by the system.
- TOTAL_OUTGOING_SMS (INT): Total number of SMS sent.
- SMS_REVENUE (Decimal): Revenue generated by SMS.
- CHARGEABLE_UNIT (INT): The amount of Units billed on SMS.
- TOTAL_DATA_VOLUME (Decimal): Amount of Data used for Internet and data related activities.
- DATA_REVENUE (Decimal): Revenue in rands, generated on Data.
- CHARGEABLE VOLUME (Decimal): Amount of Data volume billed on.
- INTERNATIONAL_PLAN (Char): Indicates if the subscriber is on International plan package or not. It can take only two values, Yes or No.
- CHURN_FLAG (Char): Indicates if the subscriber has terminated the contract or not.

The clean dataset, loaded in the database contains 24 predictors or variables and information from 5000 customers, in which the prediction is based on the field "CHURN_FLAG".

B. Min-Max Normalization, Mean, Median and Spread

In the normalization process, the distance between fields' values are adjusted to avoid the tendency of great values to influence the results of the model: scale standardization. The Minimum and Maximum values are found for every numerical predictor fields. This is to show how far the maximum value is from the minimum. The min-max normalization of a column Y, denoted Y^*_{mm} is given by:

$$Y^*_{mm} = \frac{Y - \min(Y)}{\max(Y) - \min(Y)} \quad (1)$$

The mean is used to determine the average value for every numerical field. The mean is given by:

$$\bar{y} = \frac{\sum y}{n} \quad (2)$$

Where y is the values of each filed and n is the sample size of the data, in other way the number of records in the data. The median is also calculated and used because the mean is easily affected by the outliers and noise; the median is the center of the field with the latest changed to ascending order.

The summary of the CRM data used in this paper is shown in Figure 3. For all numerical values (integers and decimals) described in section VI. A, the minimum, the maximum, the mean, the median and the Interquartile are calculated.

GROUP_SERVICE_LINE	SUBSCRIBER_STATUS_201708	REGION	INTERNATIONAL_PLAN	NATIONAL_PLAN CHURN_FLAG
Length:5899	Length:5899	Length:5899	Length:5899	No :4999
Class :character	Class :character	Class :character	Class:character	Yes: 900
Mode :character	Mode :character	Mode :character	Mode:character	
TOTAL_VOICE_DURATION	CALL_REVENUE	TOTAL_OUTGOING_SMS	SMS_REVENUE	
Min. : 0	Min. : 0	Min. : 0.00	Min. : 0.0	
1st Qu.: 5235	1st Qu.: 5296	1st Qu.: 0.00	1st Qu.: 0.0	
Median : 14788	Median : 14925	Median : 13.00	Median : 0.0	
Mean : 24846	Mean : 29399	Mean : 39.53	Mean : 818.7	
3rd Qu.: 33986	3rd Qu.: 34778	3rd Qu.: 43.00	3rd Qu.: 456.1	
Max. :363294	Max. :1155175	Max. :6403.00	Max. :213290.0	
DATA_REVENUE	CHARGEABLE_VOLUME	INTL_CALLS_REVENUE	INTL_CALLS	
Min. : 0	Min. :0.000e+00	Min. : 0	Min. : 0.00	
1st Qu.: 0	1st Qu.:5.202e+03	1st Qu.: 2812	1st Qu.: 7.00	
Median : 0	Median :5.568e+08	Median : 7749	Median : 10.00	
Mean : 8261	Mean :1.991e+09	Mean : 15583	Mean : 15.85	
3rd Qu.: 1697	3rd Qu.:2.172e+09	3rd Qu.: 18067	3rd Qu.: 18.00	
Max. :3125711	Max. :5.822e+10	Max. :1059879	Max. :523.00	
CHARGEABLE_DURATION	CHARGEABLE_UNITS	TOTAL_DATA_VOLUME	CALLS_TOTAL_NUMBER	
Min. : 0	Min. : 0.00	Min. :0.000e+00	Min. : 0.0	
1st Qu.: 4914	1st Qu.: 0.00	1st Qu.:1.257e+05	1st Qu.: 76.0	
Median : 14268	Median : 12.00	Median :5.706e+08	Median : 187.0	
Mean : 24120	Mean : 36.81	Mean :2.009e+09	Mean : 280.7	
3rd Qu.: 32857	3rd Qu.: 39.00	3rd Qu.:2.197e+09	3rd Qu.: 367.0	
Max. :330741	Max. :6403.00	Max. :5.830e+10	Max. :5016.0	

Figure 3 CRM Dataset Summary table

C. Identifying Outliers in Numerical Predictors

It is necessary to check from the dataset, values of numeric fields that are not following the trend of the rest of the data. These are outliers and can negatively affect the accuracy of the model. Certain statistical methods are sensitive to outliers [7]. Figure 4 shows the Histogram plot of International calls against their count. The graph shows the presence of no outliers, or negative numbers of international calls.

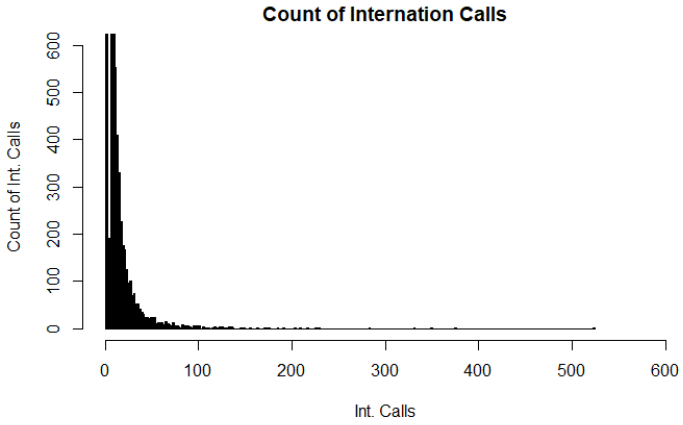

Figure 4 Histogram of International Calls count to check outliers

Most subscribers in the dataset are in the interval of 1 to 100 international calls. However, few customers have above 100 calls, which is a normal behavior in a CSP network.
From Figure 5, no abnormal values are observed. Most subscribers in the dataset also make between 0 to 1500 calls in total, both national and international. Few customers have above 1500 calls.

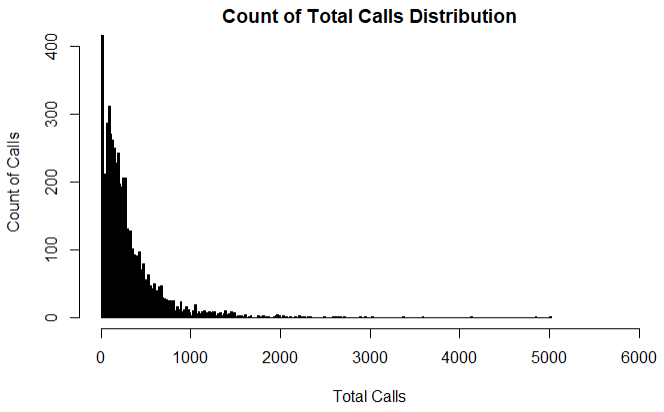

Figure 5 Histogram of Calls' count to check outliers

Most subscribers in the dataset send between 0 to 1000 SMSs, with few subscribers having more than 1000 SMSs, which is a normal behavior in the network. Illustrated in Figure 6, no abnormal phenomena or outliers are observed.

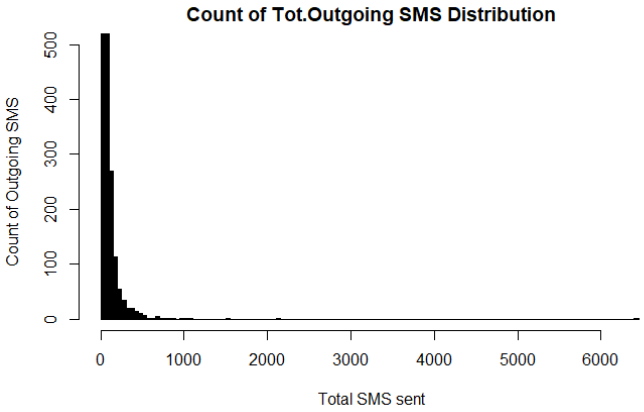

Figure 6 Histogram of SMS counts to check outliers

The relationship between the Data volume used and the chargeable data volume is shown on Figure 7.

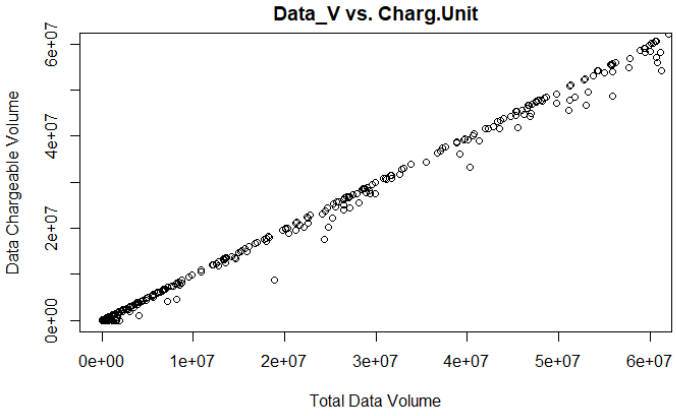

Figure 7 Graph showing the relationship between Data Volume and Chargeable Data Units

The graph shows a certain linear pattern between Data volume and Chargeable data volume. A good point to consider for the Algorithm to be used in the Predictive Analytics.

A general observation is that for and in all the predictors illustrated in the above figures, no negative or abnormal value is seen. The exposition of the patterns created by the numeric predictors provides a good insight of the customer behavior.

It also smooths the process of Machine Learning, avoiding spurious responses on the results.

VII. Exploratory Data Analysis

In this section, the relationship between predictors and the main predictor, "CHURN_FLAG" is elaborated. Considered as a continuity of Data processing, it consists to know the dataset in detail, establish the interrelationship between different variables, and observe the behavior of the variables towards the target variable.

A. Analysing Some Categorical Variables

The categorical variables on the dataset contains all the characters data type value described in section VI.A. The ultimate goal is to determine in advance from the data, the patterns that will assist scaling down the proportion of churners. Figure 8 illustrate the proportion of Subscribers who have churned vs. those who have not churned. The percentage of churned customer is 15.25%.

Table 1 Churn Proportion in the dataset

Churn Proportion	
Churn Rate	15.25682
No	Yes
4999	900

1) SERVICE_LINE:

Of type character, the field indicates the type of contract under which the customer is falling. The objective is to analyze the amount of churners per contract type.

Table 2 Churn Proportion by Service Line

Churn Flag	Hybrid	HybBB	Postpaid	Postpaid BB	Postpaid FTTH
No	2377	178	2463	31	0
Yes	298	192	302	105	3

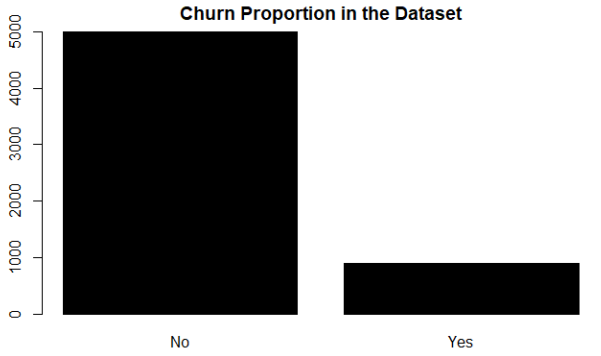

Figure 8 Churn Proportion in the CRM Dataset

2) REGION:

The objective is to analyze the amount of churners per region. Figure 9 displays the churn proportion per Region. Gauteng is the region with the highest number of churners; but also the region with the highest number of non-churners. This indicates that Gauteng is dense area.

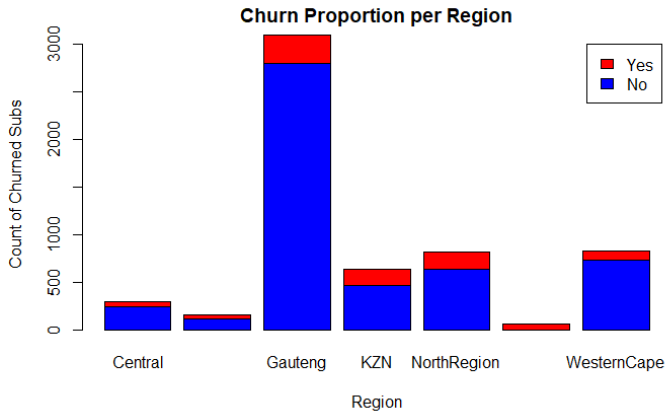

Figure 9 Subscriber churn by South African Region

Table 3 Churn Proportion by Region

Churn Flag	CENT	EC	GT	KZN	NR	UK	WC
No	246	112	2806	469	638	0	728
Yes	49	49	296	168	181	61	96

Where CENT is the Central Region, EC is Eastern Cape, GT is Gauteng, KZN is Kwazulu-Natal, NR is the Northern Region, UK is the unknown generated region and WC is the Western Cape respectively. There is no province or region in South Africa named Unknown. The CRM database is designed in a way that all transactions that have no resolved region name are classified as Unknown.

3) INTERNATIONAL_PLAN:

The objective is to analyze the amount of churners per International plan subscription. This is necessary to know in advance if most of the churners have International plan activated or not in their contract.

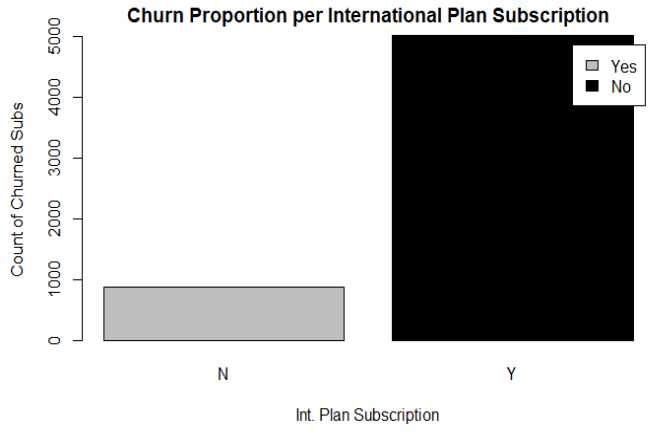

Figure 10 Subscriber Churn by International Plan Subscription

The observation from Figure 10 shows that most of the customers that have churned do not have international plan subscription. By looking at different histograms, an interrelationship can be built between the target predictor (CHURN_FLAG) and the categorical variables.

B. Analyzing Some Numerical Predictors

The objective of the numerical predictors' analysis is to dive deep into data details and uncover interrelationship between different variables and between variables and the target predictor. Using R-plotting and the result in Figure 9, the observations below are pin-pointed:

- Comparing call revenue to international call revenue, churners are in low usage, showing a trend-like or linear behavior. Illustrated in Figure 11.
- Comparing Call revenue to SMS revenue yields a similar observation, with churners showing a low revenue stream. Illustrated in Figure 12.
- Data revenue to call revenue also shows that churners show a low revenue stream on both data and calls. Illustrated in Figure 13.
- Observing the scatter plot of chargeable duration against the chargeable data volume, churners show a small chargeable units, shown in Figure 14.
- This observation will help in the building the prediction models in the sense that churners seem to be low income generating users. Illustrated in Figure 15.

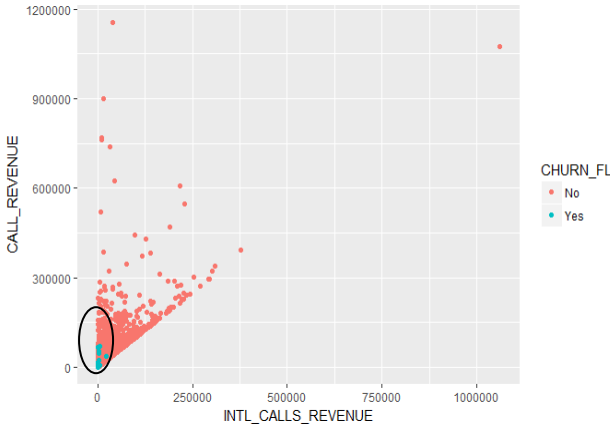

Figure 11 Call revenue vs. international call revenue

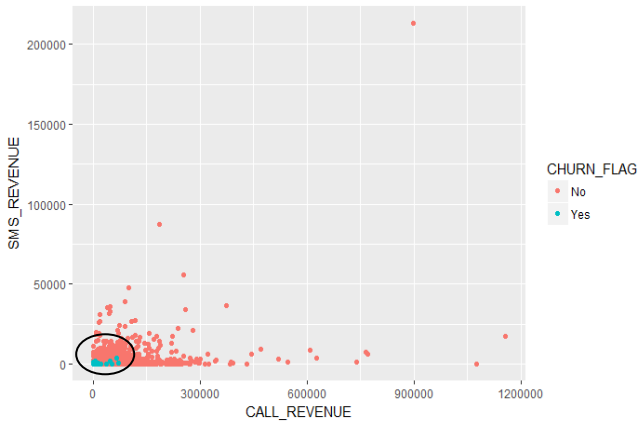

Figure 12 Call Revenue vs. SMS revenue for churners and non-churners

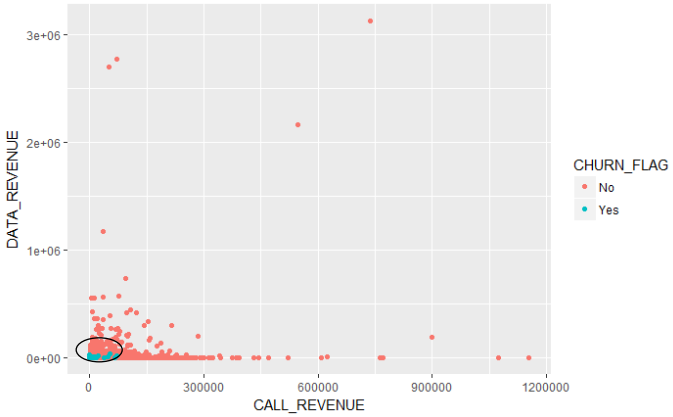

Figure 13 Data revenue vs. Call revenue for churners and non-churners

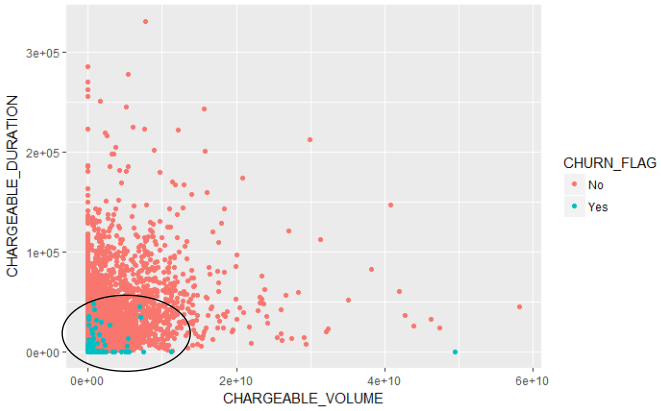

Figure 14 Chargeable duration vs. Chargeable volume of data for churners and non-churners

VIII. Machine Learning Algorithm and Model fitting

The choice of Machine Learning algorithms is not the essential part of the Predictive Analytics process. There are several machine learning algorithms built in R and Python for Predictive Analytics. As the analysis specified a target variable in the list of predictors, and the selected models have to learn the interrelationship between the target variable and the rest of predictor variables, hence supervised Machine Learning algorithms are used. The model used in this Research is tree based models, including the Classification and Regression tree which is the base of several robust algorithms such as Random Forest and Boosting models and gradient boosting models. The attractiveness of the algorithm is justified by the fact that:

- Tree based models manipulate heterogeneous datasets, containing both numerical and categorical predictor variables.
- Tree based models are resistant to noise and outliers, to some extent.
- Tree based models are interpreted easily, independent of the knowledge of the users in Data Science and Statistics.
- Tree based models easily identifies hidden relationships between predictor variables, regardless of the complexity.

An Illustration of a Tree process model is shown in Figure 15.

A. Data Partitioning and Model Fitting

The processed or cleaned dataset is split into two sub datasets, training and testing sets. The partition is such that 60% of the dataset is used to train the model and 40% is used to test the model. Table 4 shows the dimension of the split datasets.

- *inTrain=createDataPartition(y=db_frame$CHURN_FLAG, p=0.7, list = FALSE)*
- *db_training=db_frame[inTrain,]*
- *db_testing=db_frame[-inTrain,]*

Where inTrain is the partitioned dataset variable, db_frame is the main processed dataset, CHURN_FLAG is the target variable, db_training is the training dataset and db_testing is the testing dataset. Figure 14 shows the dimensions of the training dataset and testing dataset after partition.

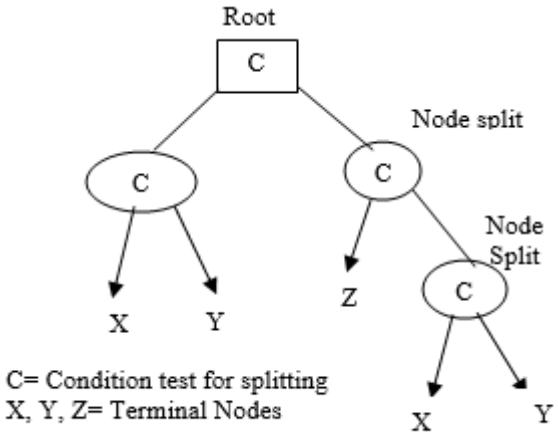

Figure 15 Decision Tree Process Model

Table 4 Dimension of training and testing datasets

	Number of Records	Number of Predictors
training dataset:	4130	17
Testing dataset:	1769	17

B. Fitting the Model

- *model_lm=train(CHURN_FLAG~.,method="rpart", data=db_training)*
- *model_lm*

```
CART

4130 samples
  16 predictor
   2 classes: 'No', 'Yes'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 4130, 4130, 4130, 4130, 4130, 4130, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.0000000  0.9991811  0.9967795
  0.4992063  0.9991811  0.9967795
  0.9984127  0.9500992  0.6772862

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was cp = 0.4992063.
n= 4130

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 4130 630 No (0.8474576271 0.1525423729)
  2) INTL_CALLS>=5.5 3501   1 No (0.9997143673 0.0002856327) *
  3) INTL_CALLS< 5.5 629   0 Yes (0.0000000000 1.0000000000) *
```

Figure 16 CRM training set Model Fit with regression and classification trees

Based on the model fit's output result, the classification and Regression tree model provides an accuracy of 99.918%.

C. Gradient Boosting Model and Random Forest

Gradient boosting algorithms are used to optimize the prediction accuracy and using gradient descent techniques, statistical model estimates are obtained. An important point on the gradient boosting is that variable selection is conducted during the model fitting process [8], with no dependency on heuristic methods such as stepwise selection.

For an outcome variable *y*, and a number of predictor variables $x_1, x_2, \ldots x_p$, the objective is to model the relationship between *y* and $x := (x_1, x_2, \ldots x_p)^T$ and achieve an optimal prediction of *y* provided *x*. The above requirement is achieved by minimizing the loss function $\rho(y, f) \in \mathbb{R}$ over the function *f*. Thus, in the gradient boosting environment, the ultimate goal is to estimate the optimal prediction function f^* given by [9]:

$$f^* := \operatorname{argmin}_f \mathbb{E}_{Y,\mathbf{X}}\left[\rho(y, f(\mathbf{x}^\top))\right], \qquad (3)$$

Since (1) is a predictive function, the expectation is not known. Instead of minimizing the loss function $\rho(y, f) \in \mathbb{R}$, boosting models minimize the mean, *R* given by:

$$\mathcal{R} := \sum_{i=1}^{n} \rho(y_i, f(\mathbf{x}_i^\top)) \qquad (4)$$

In this paper, the two methods used to optimize the prediction are boosting trees and Random Forest. In the R environment, these algorithms are provided by the function "gbm" and "rf" respectively.

1) Boosting Trees

Boosting Trees reduces bias and variance in supervised Machine Learning [10]. As one of the gradient boosting algorithms, it is based on weak learners, predictor variables which highly biased and lowly variant.

The model is trained by fitting the training dataset, db_training into it. The result of the model fit is shown in Figure 19.

- *model_bt=train(CHURN_FLAG~., method="gbm", data = db_training, verbose=FALSE)*
- *model_bt*

```
4130 samples
  16 predictor
   2 classes: 'No', 'Yes'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 4130, 4130, 4130, 4130, 4130, 4130, ...
Resampling results across tuning parameters:

  interaction.depth  n.trees  Accuracy   Kappa
  1                   50      0.9993463  0.9974436
  1                  100      0.9992664  0.9971374
  1                  150      0.9992928  0.9972350
  2                   50      0.9993728  0.9975490
  2                  100      0.9992928  0.9972350
  2                  150      0.9992664  0.9971388
  3                   50      0.9993728  0.9975490
  3                  100      0.9993194  0.9973401
  3                  150      0.9992928  0.9972350

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning parameter 'n.minobsinnode' was held constant at a value of 10
Accuracy was used to select the optimal model using  the largest value.
The final values used for the model were n.trees = 50, interaction.depth = 2,
shrinkage = 0.1 and n.minobsinnode = 10.
```

Figure 17 Prediction Model fit with gradient boosting

The best achieved accuracy with boosting tree algorithm is 99.937%.

2) Random Forest

Random Forest is a more recent supervised Machine Learning algorithm built on regression tree, developed by L. Brieman [11], to optimize the prediction accuracy without overfitting the data. The predictor variables are ranked in an unbiased way comparing to other regression models, which provides the best accuracy [12].

Using the training set db_training, the Random Forest algorithm is used to train and fit the model as below. The result of the model fit is shown in Figure 20.

- *model_rf=train(CHURN_FLAG~.,method="rf", data = db_training, prox=TRUE)*
- *model_rf*

```
Random Forest

4130 samples
  16 predictor
   2 classes: 'No', 'Yes'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 4130, 4130, 4130, 4130, 4130, 4130, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.9993399  0.9974455
  13    0.9994970  0.9980540
  25    0.9994448  0.9978504

Accuracy was used to select the optimal model using  the largest value.
The final value used for the model was mtry = 13.
```

Figure 18 Prediction Model fit with Random Forest Algorithm

The best achieved accuracy with Random Forest is 99.949%.

IX. Model Evaluation

Three Machine Learning algorithms have been used to train the CRM dataset including the classification tree, Boosting trees and Random Forest. The following comparison is deduced from the result:

Table 5 Model Evaluation Result per Algorithm

	Machine Learning Algorithm Evaluation		
Evaluation parameter	Classification Tree	Boosting Trees	Random Forest
Accuracy	99.918 %	99.937 %	99.949 %
In-Sample Error	0.082	0.063	0.051

1) Predicting new Data

Random Forest is the model that has given the best accuracy, with an in-sample error of 0.051. The model is used for the prediction of new dataset (testing set). However, an evaluation with classification tree is shown in terms of new dataset.

- *newdataset=db_testing[,-c(17)]*
- *predictions=predict(model_rf, newdata = newdataset)*
- *confusionMatrix(predictions, db_testing$CHURN_FLAG)*

Figure 19 shows the output result of the Random Forest model built and trained for the purpose of churn prediction. The model has predicted to 100% accuracy in this case.

In Figure 20, the classification tree method is shown with the testing site. And the accuracy can be seen at 99.89% with 2 false positive results.

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1499    0
       Yes    0  270

               Accuracy : 1
                 95% CI : (0.9979, 1)
    No Information Rate : 0.8474
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.8474
         Detection Rate : 0.8474
   Detection Prevalence : 0.8474
      Balanced Accuracy : 1.0000

       'Positive' Class : No
```

Figure 19 Prediction Performance on new Data using Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1499    2
       Yes    0  268

               Accuracy : 0.9989
                 95% CI : (0.9959, 0.9999)
    No Information Rate : 0.8474
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9956
 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 1.0000
            Specificity : 0.9926
         Pos Pred Value : 0.9987
         Neg Pred Value : 1.0000
             Prevalence : 0.8474
         Detection Rate : 0.8474
   Detection Prevalence : 0.8485
      Balanced Accuracy : 0.9963

       'Positive' Class : No
```

Figure 20 Predictive Performance on new Data using Classification Tree

2) The Confusion Matrix

The **confusion matrix** is a great indicator of the Model performance, a table that classifies predictions according to whether they match the actual value or not. One of the table's dimensions indicates the possible categories of predicted values, while the other dimension indicates the same for actual values. If the predicted value matches the actual value, then the model provides an exact classification. Thus, the matrix is made of Positive and negative classes of predicted values. The correct predictions are named "True" which can be either True positive or True Negative. Based on the model performance above, the confusion Matrix is explained as follow in Table 6:

Table 6 Confusion Matrix

		Predicted Churn	
Actual Churn		No	Yes
	No	True Negative (Correctly classified as Non-churners)	False Positive (Incorrectly classified as Churners)
	Yes	False Negative (Incorrectly classified as Non-churners)	True Positive (Correctly classified as Churners)

Note that the Accuracy and the sample error rate are calculated using the Confusion Matrix where Accuracy A is given by:

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

And the Sample Error is given by: $Err = 1 - A$

Where TP = True Positive, TN=True Negative, FP=False Positive and FN=False Negative.

3) Precision and Recall

Calculating the precision and recall give a great indication of performance with accent on how relevant the results of the model is. The effect of noise in the prediction function is determined by the Precision and Recall. The precision is the proportion of positive predicted values, and the Recall is the measure of precision of the result.

$$Pr = \frac{TP}{TP+FP} \quad \Longrightarrow \quad \text{Pr} = 1$$

$$Rec = \frac{TP}{TP+FN} \quad \Longrightarrow \quad \text{Rec} = 1$$

The precision and Recall of 1 provide an excellent result of the model.

4) Predictor Variable Importance

The Random Forest model has provided the best performance. It is also important to determine the predictor importance by categorizing the effect of the predictors on the model. From R the generic function ***varImp*** is used to work out the variable importance.

Table 7 Predictor Importance

Predictor	Overall
INTL_CALLS	445.8
SUBSCRIBER_STATUS_201708Suspended	362.3
INTL_CALLS_REVENUE	146.3
INTERNATIONAL_PLANY	80.01
CALL_REVENUE	23.09
CHARGEABLE_DURATION	6.032
TOTAL_VOICE_DURATION	4.022
CHARGEABLE_UNITS	0.1515
DATA_REVENUE	0.141
TOTAL_OUTGOING_SMS	0.1188
CALLS_TOTAL_NUMBER	0.07219
CHARGEABLE_VOLUME	0.06375
TOTAL_DATA_VOLUME	0.03731
SMS_REVENUE	0.02417
REGIONWesternCape	0.01914
GROUP_SERVICE_LINEPostpaid	0.0084
GROUP_SERVICE_LINEPostpaidBroadband	0.006667
REGIONEasternCape	0.004756
REGIONGauteng	0.002667
GROUP_SERVICE_LINEHybridBroadband	0

5) Displaying the Predicted Data

The customers that have been identified, likely to churn in the future can be displayed using the data frame joining function. The false positive predictions for classification tree model is also be viewed in the dataset by adding an additional column of the predicted results to the original new data set.

- *prdictium=data.frame(db_testing,test_pred)*
- *predicted_churners=data.frame(db_testing, predictions).*

X. Conclusion

One of the fundamental points in Data Analysis and Machine Learning is the development and choice of the algorithm for building the model and selecting the correct variables [13]. However, doing a predictive Analytics is more than just choosing a Machine Learning algorithm to predict. A thorough understanding of the dataset and variables is mandatory, especially in the case of supervised Machine Learning. The choice of the algorithm depends on many parameters where in this case, only the accuracy and the sampling error have been taken into consideration.

The Out-of-sample error, which is the error rate on the new data set is the indicators that counts the most. On the new dataset, Random Forest provides an out-of-sample error of "0" and the Classification Tree provides an out of sample error of "0.12" which is higher than the in-sample error rate (sampling error on the training site). For this reason, classification tree is not very suited to use in this case, as it is preferable to have the out-of-sample error lower than the in-sample-error.

It is obvious that we need a high accuracy on the new dataset where decisions and future steps are made.

In summary, the below performance is achieved on the case study:

- Random Forest: The model with the best performance: 0.051 > 0 (sampling error). Good accuracy on predicting new dataset, with the precision and recall of 1.

Table 8 Prediction Summary for Random Forest

	On the training set	On the testing set (new dataset)
Accuracy	99.949	100
In-sample error, Out-of sample error	0.051	0

XI. Future studies

The area of Data Science, Big Data and Machine Learning are expanding and growing at a fast pace. Such is the area of Telecommunications with the implementation of 5G and the rise of the Internet of Things (IoT). CSPs (Communication Service Providers) will need to adapt their infrastructure to embrace new technologies. That includes integrating Predictive Analytics and Machine Learning in to their business strategy to stay ahead of competition.

The more data, the better for the Algorithms to learn. The CRM platform contains only customer related information. In order to take advantage of Machine Learning and Predictive Analytics, the scope needs to be broaden to accommodate multiple cases. In the future, Machine Learning has to integrate all different sources of CSP data including QoS (Quality of Service) data. Machine Learning Algorithms such as Neural Networks, which provide an improved performance on accuracy and learning, comparing to other traditional models need to be considered.

REFERENCES

[1] A. Azzalini and B. Scarpa, Data Analysis & Data Mining, and Introduction, Oxford University Press, 2012, pp 2-5.

[2] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[3] Eric Siegel, Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. 1st. Wiley Publishing ©2013 ISBN:1118356853 9781118356852, pp 10-13.

[4] Eric Siegel, Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. 1st. Wiley Publishing ©2013 ISBN:1118356853 9781118356852, pp 10-13.

[5] A. Azzalini and B. Scarpa, Data Analysis & Data Mining, and Introduction, Oxford University Press, 2012, pp 4.

[6] Mayer-Sch€onberger, V., and Cukier, K. 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think. New York: Houghton Mifflin Harcourt Publishing Company.

[7] Daniel T. Larose and Chantal D., Data Mining and Predictive Analytics, 1st Edition, Larose. ©2005 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

[8] P. Buhlmann and B. Yu. Boosting with the¨ L2 loss: Regression and classification. Journal of the American Statistical Association, 98:324–338, 2003.

[9] Jake Morgan, Classification and Regression Trees Analysis, Boston University school of Public Health, Technical Report No. 1, May 8, 2014

[10] Leo Breiman (1996). "BIAS, VARIANCE, AND ARCING CLASSIFIERS" (PDF). TECHNICAL REPORT. Archived from the original (PDF) on 2018-01-19. Retrieved 19 January 2018. Arcing [Boosting] is more successful than bagging in variance reduction.

[11] L. Breiman. Random forests. Machine Learning, 45:5–32, 2001.

[12] Camino Gonzalez, José Mira-McWilliams1, Isabel Juárez2, Important variable assessment and electricity price forecasting based on regression tree models: classification and regression trees, Bagging and Random Forests, Statistical Laboratory, Escuela Técnica Superior de Ingenieros Industriales, Technical University of Madrid, c/José Gutiérrez Abascal, 2, 28006 Madrid, Spain, ISSN 1751-8687.

[13] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, 2 edition, 2009.

[14] James Manyika *et al.*, Mckinsey, Big data: The next frontier for innovation, competition, and productivity, by Global Institute, www.mckinsey.com, May, 2011. Last accessed March 16, 2014.

[15] Peter Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinart, Colin Shearer, Rudiger Wirth, CRISP-DM Step-by-Step Data mining guide, 2000

[16] Matthew A. Waller and Stanley E. Fawcett, Data Science, Predictive Analytics, and Big Data: A Revolution

[17] That Will Transform Supply Chain Design and Management, Journal of Business Logistics, 2013, 34(2): 77–84 © Council of Supply Chain Management Professionals.

[18] P. Buhlmann and T. Hothorn. Boosting algorithms: Regularization, prediction and model fitting (with¨ discussion). *Statistical Science*, 22:477–522, 2007.

[19] Benjamin Hofner, Andreas Mayr, Nikolay Robinzonov and Mattthias Schmid (2014), Mo del-based Bo osting in R – A Hands-on Tutorial Using the R Package mboost. *Computational Statistics*, 29:3–35.