Real time customer churn scoring model for the telecommunications industry

Nyashadzashe Tamuka
Department of Computer Science,
University of Fort Hare
Private Bag X1314 Alice 5700, South Africa
tnyashadzashe@gmail.com

Khulumani Sibanda
Department of Computer Science,
University of Fort Hare
Private Bag X1314 Alice 5700, South Africa
ksibanda@ufh.ac.za

2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC) | 978-1-7281-9520-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/IMITEC50163.2020.9334129

***Abstract*—There are two types of customers in the telecommunication industry; the pre-paid and the contract customers. In South Africa it is the pre-paid customers that keep telcos constantly worried because such customers do not have anything binding them to the company, they can leave and join a competitor at any time. To retain such customers, telcos need to customise suitable solutions especially for those customers that are agitating and can churn at any time. This needs customer churn prediction models that would take advantage of big data analytics and provide the telco industry with a real time solution. The purpose of this study was to develop a real time customer churn prediction model. The study used the CRISP-DM methodology and the three machine learning algorithms for implementation. Watson Studio software was used for the model prototype deployment. The study used the confusion matrix to unpack a number of performance measures. The results showed that all the models had some degree of misclassification, however the misclassification rate of the Logistic Regression was very minimal (2.2%) as differentiated from the Random Forest and the Decision Tree, which had misclassification rates of 20.8% and 21.7% respectively. The results further showed that both Random Forest and the Decision Tree had good accuracy rates of 78.3% and 79.2% respectively, although they were still not better than that of the Logistic Regression. Despite the two having good accuracy rates, they had the highest rates of misclassification of class events. The conclusion we drew from this was that, accuracy is not a dependable measure for determining model performance.**

Keywords— machine-learning, churn, unbalanced classes, telcos, continuously updated, real time / instant churn scoring, continuous learning, performance monitoring.

I. INTRODUCTION

Telecommunication industry is the backbone of any nation's communication infrastructure. More so in the era of industrial revolution where everything is destined to be digitalized. Many sectors now depend on telecommunication industries for their day to day operations. For example, the banking sector, the retail sector, the manufacturing sector, the health sector, and all individual citizens now rely on telecommunications infrastructure for their business operations and social interactions. This underlines the importance of telecommunication industries to any nation's economic growth and also to promote social cohesion, in the case of individual citizens. The industry contributes massively to the national gross domestic product (GDP) through mobile and fixed communication networks. It also provides employment to millions of citizens whose livelihoods and their dependents are depended on the industry. However, for its growth and sustainability, the industry is emphatically dependent on its customers. Luckily for the industry, the customer base has been increasing exponentially as a result of Internet services demand and other services like voice services. The sector has not only seen the increase of customers but also an increase in the number of telecommunication companies. This has triggered a fierce competition for the customers, leading to a number of business strategies to attract new and retain existing customers.

Telecommunication industry customers have two segments; the subscribers and the pre-paid customers. Pre-paid customers consist the largest percentage for the total number of customers. For example, the South African had a total of 96 million mobile voice subscribers in 2019 [1]. Of those, 82 million were pre-paid subscribers while 14 million were contract customers. It is mostly the pre-paid segment of customers that keep the telecommunication companies fighting for survival. This segment of customers is very jittery, customer can churn at any time, in other words they can leave one company and join another at any time since they are not tied down by any contract. Customer churn occurs when customers quit subscribing to a company in preference to another company [2]. This results in the loss of company's customers which leads to increased costs and reduced revenue. Considering that pre-paid customers consist the largest segment, battle lines have been drawn amongst the telecommunication companies, each fighting to keep its own customers while simultaneously pursuing other companies' customers. This is a very taxing exercise that needs companies to implement complex business strategies for competitive leverage. This would also require using accurate models for identifying subscribers that are prone to agitate and join a rivalry company.

Customer churn is a serious problem to telecommunication companies and those that are not able to retain their subscribers, would face limited sales, increased company costs, deteriorating profit margins and bad company image. In order to detect possible customer churners, companies would need to make use of prediction models that are not only efficient but also real time adaptive. This research aimed at developing and testing a model that can detect telco customers who are likely to churn in real time, when continuously updated using new subscriber records. Several researchers have proposed similar studies that acted as references for this study. Most of these studies have implemented machine learning algorithms and reported varying degrees of accuracies. In the following section we review a number of the related studies on subscriber churn prediction.

A. Related studies

The Random Forest algorithm was implemented by [3] to predict subscriber churn. Their study revealed that the characteristics of big data (velocity, veracity volume and variety) gathered by telcos, considerably improves customer churn prediction. The authors used Apache Hadoop to store big data with large set of features/attributes. They gathered the

data from the business support and operations support sectors at one of the largest Chinese mobile company. The researchers used Apache Spark to process huge volumes of data. They evaluated the model using the ROC curve and results indicated that the model was accurate; however, their focus was on batch processing of historical data. The weakness of their approach is that it is difficult to update the model to predict using continuously generated big data streams.

A related study to [3] was conducted by [4]. The authors applied four machine-learning algorithms on a big data platform to develop an accurate churn identification model. They used big data provided by SyriaTel, a Syrian telecommunications company. In their model evaluation, the XGBOOST algorithm outperformed other algorithms by producing the most accurate performance of 93.3%. The model performed well in batch scoring of historical customer data; nevertheless, it is not possible to obtain real time predictions using the model on new data.

A neural network was applied to predict customer attrition in a big telecommunications company in China [5]. The authors further divided the churn customers predicted by the RBF neural network into four groups using the Analog Complexion algorithm, so that telcos could implement suitable actions to retain the customers. Although the overall accuracy rate obtained was 91.1%, again it was impossible to update the models with constantly changing features required for real time customer churn prediction.

[6] contributed to the notion of churn prediction by proposing a new framework for a churn prediction model and the authors implemented the framework using WEKA software. They categorized subscribers as possible churners or non-churners. They trained logistic regression and decision tree algorithms with three various datasets which were small, medium and large in sizes. They compared the performance of the two algorithms, and the decision tree was more accurate than logistic regression. Their framework was suitable to develop a churn prediction model, however only batch scoring using offline customer data is viable since the authors did not present how their model can identify churners when using constantly changing customer attributes.

A methodology to predict subscriber churn in telcos using a dataset with 21 features for 3333 prepaid customers' call details was presented by [7]. The dataset consisted of a dependent attribute (variable) with two churn categories: Yes / No. They implemented Principle Component Analysis (PCA) for dimensionality reduction by removing correlated features. They applied three machine-learning algorithms to identify customer churn. The ROC evaluated the algorithms' performance and the Bayes Networks produced the highest accuracy. The main drawback of their study was that the model could not be updated with new data.

As indicated in the few works that have been reviewed, most models are built and adapted for batch datasets (already existing customer data) without factoring in new data that would keep changing the equation of churn and non-churner customers. The proposed models are batch system based, that is, they are limited when it comes to continuously changing telcos datasets. The desire to build a customer churn model that can be continuously updated for instant churn prediction / scoring was the motivator for this study. The model was implemented in the big data environment (i.e. Apache Spark from IBM Watson Studio). The Watson Studio is available on IBM cloud [8]. The rest of this paper is organized as follows: section II discusses on this study's methodology and methods, the results are presented in section III, section IV discusses the results. This study concludes in section V

II. METHODOLOGY

This study aimed at building an accurate model that can predict subscriber churn instantly (real time churn scoring) when continuously supplied with new subscriber records. The study required a good understanding of the business case which should culminate to a suitable model deployment. Within the ambit of the domain of this study, there are generally two methodologies that could be used; the SEMMA and the CRISP-DM (Cross Industry Standard Process for Data Mining) methodologies [9]. The SEMMA consists of five sequential phases that guide the implementation of projects in data science. Its phases involve "Sample, Explore, Modify, Model and Asses" [10]. For the purposes of this study, the CRISP-DM methodology was adopted because of its ability to guide understanding of the business case. The SEMMA places less emphasis on the initial planning phases of a project, something that can be detrimental when developing predictive models as their accuracy depends on proper understanding of the business case. The CRISP-DM methodology consists of six various stages which are presented in Fig. 1.

Fig. 1. *The CRISP-DM*

A. The design of the study

The design constitutes the overall methods for carrying out the project implementation. For this study the whole strategy was guided by the phases of the methodology adopted. It is those phases that constituted the plan for data gathering, data cleaning and data analysis. As shown in Fig. 1., the six phases are discussed in the following subsections.

1) Business understanding

Understanding the business' requirements / needs includes setting business objectives and defining a plan to solve the problem. This is of paramount importance in identifying churn. In this phase, the task of identifying churners is based on the telcos business perspective. The required resources like dataset sources were identified and the first assessment of the required software that would be necessary was made.

2) Data Understanding

Data understanding involves the process of collecting data, data description and data exploration. The following subsections give details of the whole process we undertook to understand the data used for the study.

a) Data Collection

For this study, we used secondary data to avoid a lot of formidable hurdles that could be associated with obtaining clients data of a telecommunications company. We freely obtained secondary data from Kaggle [11]. Kaggle is a global data science organization with thousands of datasets that can be used for machine learning projects and available for free. We extracted the dataset from the Kaggle website and then began the process of scrutinizing the data.

b) Data description

The sample dataset consists of 7043 unique customers who subscribed to a telecommunication company called Telco. The dataset consists of 7043 rows and 21 columns. Rows contain the customers' records. Columns consists of numerical and categorical attributes.

c) Data Exploration

Data exploration was conducted using Spark's Dataframe *methods.printSchema()* to print the dataset in a tree format. Collected data was scrutinized for useful features, we also evaluated the dataset to determine if it would meet the study objectives. A scatterplot matrix was applied for correlation analysis to identify the relationships between the independent numerical attributes in the dataset. The correlation values vary from -1 to 1. A value near -1 highlighted a strong negative relationship (correlation). A value near 1 highlighted a strong positive relationship (correlation). Highly correlated independent attributes reduce the model's accuracy since it is difficult to assess their contribution to the dependent attribute [12]. We also checked for missing values and errors in the dataset to ensure data integrity.

3) Data preparation / pre-processing

The performance of the churn prediction model depends on data integrity. This is the most important phase and it is time consuming. It consists of all the activities required to transform the raw Telco sample dataset into the dataset that was suitable for modelling. For this study, it included removing missing values, dropping attributes, handling imbalanced / unbalanced classes, transforming attributes (encoding) and splitting the dataset.

a) Handling missing values

Most algorithms perform poorly when the training dataset contains missing values. We found that the dataset had some attributes that contained missed values. This could be because the subscribers did not provide the required fields. As a rule of thumb, attributes that have more than 60% of values missing are normally dropped [13]. For this study, the mode which occurred frequently in the attributes was utilized to impute the values missing, this was in line with other researchers' procedures like [14]. In particular, numerical attributes with values missing which ranged from 2% to 30% were imputed by mean values as recommended in [14]. The premise to this procedure is that, mean is an acceptable approximation for the values selected from a normal distribution.

b) Dropping irrelevant attributes

Some attributes were found not to be useful for training the algorithms. Such attributes were dropped since they were not essential for churn prediction. We checked for such attributes and identified the CustomerID attribute, and dropped it from the Telco dataset.

c) Feature selection

This was implemented to select the best features / attributes for developing the subscriber churn prediction model. The selection of best features was implemented based on the correlation of the Telco dataset's independent attributes presented by the heat map of the attributes' correlation. The aim was to drop the two highly correlated attributes which was less correlated to the Churn attribute [15]. To assess the contribution of the independent attributes to the Churn attribute, it was mandatory to filter highly correlated features. This was necessary to improve the models' accuracy and to train the algorithms faster.

d) Transforming attributes

Some algorithms can process datasets of specific data types, for example, logistic regression cannot process categorical attributes which are in a string format like Yes or No. The categorical attributes in the dataset were encoded to 1 or 0. Various techniques can be implemented to transform attributes' data types. In Watson studio the attributes were encoded automatically during model building. The Telco dataset's attributes are presented in TABLE 1.

TABLE 1. THE TELCO DATASET'S ATTRIBUTES

Attribute name	Description	Data type
customer ID	Unique ID for each subscriber	String
Gender	The subscriber's gender (i.e. Male / Female)	String
SeniorCitizen	The subscriber is a senior citizen or not (1 / 0)	Numeric
Partner	The subscriber has a partner or not (Yes / No)	String
Dependents	The subscriber has dependents or not (Yes / No)	String
Tenure	Number of months subscribing to the company	Numeric
PhoneService	The subscriber has phone service or not (Yes / No)	String
MultipleLines	The subscriber has multiple service or not (Yes / No / No Phone)	String
InternetService	Subscriber's internet service provider (DSL / Fibre Optic / No)	String
OnlineSecurity	The subscriber has online security or not (Yes / No / No internet service)	String
OnlineBackup	The subscriber has online backup or not (Yes / No / No internet service)	String
DeviceProtection	The subscriber has device protection or not (Yes / No / No internet service)	String
TechSupport	The subscriber has tech support or not (Yes / No / No internet service)	String
StreamingTV	The subscriber has streaming service / not (Yes / No / No internet service)	String
StreamingMovies	The subscriber has streaming movies service or not (Yes / No)	String
Contract	The contract type for a subscriber (i.e. Month-to-month, One year, Two year)	String
PaperlessBilling	The subscriber has paperless billing or not (Yes / No)	String
PaymentMethod	Payment method (i.e. Bank Transfer, Electronic Check, Credit Card, Mailed Check)	String
MonthlyCharges	Monthly charges for the customer	Numeric
TotalCharges	Total charges for the customer	Numeric
Churn	Did the subscriber leave the company or stay	String

e) Unbalanced classes

The distribution of the target attribute was visualized as shown in Fig. 2. The not Churn class is represented by class 0 and the Churn class is represented by class 1 in the Figure. The Figure shows 1869 churners and 5174 non-churners, implying that the target attribute was not equally distributed. The churn attribute didn't contain equal amounts of churners and non-churners; hence it was unbalanced (unequally distributed). The non-churners were 76% of the total subscribers and the churners were 24% of the total subscribers. The dataset was imbalanced because the not Churn class / 0 was more dominant than the Churn class / 1. The task was now to apply a method that could help to equally distribute the classes.

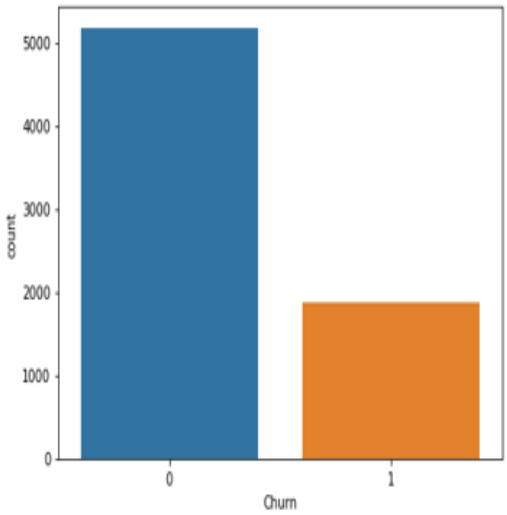

Fig. 2. *The distribution of the target variable*

The algorithm level and data level methods are important in balancing the classes. In a study by [16], these methods use up-sampling and down-sampling which are data level methods to balance the classes. These methods have some disadvantages for example important records can be dropped from the dataset by down-sampling and up-sampling can cause overfitting. In this study, the cost-sensitive approach was adopted to equally distribute the classes. A big weight was assigned to the minority class and the small weight to the majority class to deal with unbalanced classes [17]. The cost sensitive method was applied because it enables the use of all the dataset without sampling, hence avoiding the loss of important records when training the algorithm.

f) Dataset splitting

It was important to split the dataset into training dataset which was used for building the models from the algorithms (i.e. learning of the algorithms) and the testing dataset for evaluating the models' performance. The Telco sample dataset was randomly split into 80% training and 20% testing datasets to avoid the bias of the models [18].

4) Modelling

For this study, this phase involved selecting the algorithms to be used for our implementation. We therefore selected three algorithms to train and compare their performances to choose the best performing one. The algorithms that were trained and evaluated are: Logistic Regression, Random Forests and Decision Tree. These algorithms were selected after carefully studying literature which showed that they tend to perform better in real time churn scoring [19]. The modeler flow in Watson studio contained the algorithms which were trained with the already prepared data to develop the models. We started by added the project to the Watson studio modeler flow and named it 'Churn prediction', after which Spark environment was activated. The pre-processed dataset was then imported. Machine learning algorithms were selected from the modeler environment and the dataset was then linked to the algorithms. The next task was then to select the target and the independent features. At that moment, the algorithms were then trained using the train dataset. After developing the models, we then described the resulted models and documented the interpretation of the resulting models.

5) Evaluation

Evaluation involves testing the model with unseen data and checking if the model can meet objectives. During this phase we assessed the degree at which the resultant models were meeting the study objectives. One of the commonly used model evaluation techniques is the confusion matrix.

It is through the confusion matrix that all other evaluation measures can be calculated. The following measures were calculated: accuracy, recall, F-measure and precision A sample of 3498 subscribers was used for evaluating the model. The performance of the models was compared using learning curves and lift charts. The best model was chosen for deployment. Fig. 3 presents a confusion matrix.

	Actual positive	Actual Negative
Predicted positive	True Positive (TP)	False Positive (FP)
Predicted negative	False Negative (FN)	True Negative (TN)

Fig. 3. *The confusion matrix*

TP / True Positives: the number of subscribers who churned and the model correctly classifies them as churners.

TN / True Negatives: the number of subscribers who didn't actually churn and the model classifies them as non-churners.

FP / False Positives: the number of subscribers who didn't actually churn and the model classifies them as churners.

FN / False Negatives: the number of subscribers who actually churned and the model classifies them as non-churners.

a) Metrics calculated from the confusion matrix

Precision: The proportion of total churners predicted that actually churns. The best model was the one which yielded the highest precision.

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

Recall: The proportion of actual churners who were correctly classified as churners. The best model was the one which produced the highest recall for churners.

$$\text{Recall} = \frac{TP}{(TP+FN)}$$

F-measure (F_1 score): The harmonic average of recall and precision (Fawcett, 2006). Precision and recall were calculated from the confusion matrix. This was the best metric to evaluate the models since they were classifying the imbalanced Telco dataset. The F-measure ranges from 0 to 1, with highest value e.g. 0.9 indicating outstanding model's performance and a lowest value e.g. 0.2 showing poor performance.

$$F_1 \text{ score} = \frac{2*Precision*Recall}{(Precision+Recall)}$$

Accuracy: the number of churners or non-churners who are correctly predicted over total number of subscribers.

$$\text{Accuracy} = \frac{TP+TN}{(TP+TN+FP+FN)}$$

6) Model deployment

There are three options for deploying the model in Watson studio, through the web service, batch prediction or real-time streaming predictions. Web service enables the integration of

the model with a mobile or web application. Real-time streaming predictions enables predictions on continuously generated data streams. The model was tested by dynamically providing input attribute values for one customer at a time and churn prediction scores provided in immediately. The new data required for continuous learning was stored in the DB2 warehouse on IBM cloud. After evaluating the models, logistic regression was selected for deployment as it outperformed other models. The model was deployed as a web application and it used the API to analyze new customer records loaded and perform real time (instant) predictions for individual customers. The new attribute values were absent from the Telco sample training dataset. Churn scores / probability were produced in real time. The DB2 warehouse is a distributed elastic data warehouse service developed by IBM [20]. Continuous learning in Watson studio allows automatic performance monitoring of the models using evaluation metrics, retraining of the models and redeployment to maintain the models' performance [21].

a) Continuous learning

The Performance Monitoring function of the Watson Studio automatically updated the algorithms with new data available (continuous learning), evaluate and retrain the algorithms when performance is below the required threshold, hence constantly improving model's accuracy.

III. RESULTS

In TABLE 2 we can note that the logistic regression-based model had a recall and accuracy of 97.8%, while the decision tree had 78.3% accuracy and recall. The random forest had an accuracy and recall of 79.2%. This shows that in terms of accuracy and recall, the logistic regression fared better than the other two models. Having a better accuracy and better recall is an indication that, for weighted class 0 and class 1, the logistic model correctly classified the data instances more than the other two models. Furthermore, TABLE 2. shows that the logistic regression had a false positive rate of 5.4%, while the decision tree and the random forest had 35.7% and 40.9% respectively. Again, this shows that the logistic regression had the least probability of wrongly misclassifying non-churners as churners. In the same vein, the F1 measure of 97.8% shows that the logistic model produced an outstanding performance since it was classifying more of the minority category (i.e. churn). On the part of the Random Forest, a recall of 79.2%, precision of 78%, accuracy of 79.2% and the F1 measure of 77.8% shows that the Random Forest model performed better than the Decision Tree-based model even though it provided less information about the churners. For the Decision Tree, a recall of 78.8% and the precision of 77.7% shows that the model was misclassifying many churners. Finally, the comparison of the models shown in TABLE 2. presented Logistic Regression model as the best model for subscriber churn prediction.

TABLE 2. THE SUMMARY OF THE EVALUATION METRICS

	Logistic Regression model	**Decision Tree Model**	**Random Forest Model**
Accuracy	97.8%	78.3%	79.2%
True Positive Rate	97.8%	78.3%	79.2%
False Positive Rate	5.4%	35.7%	40.9%
Precision	97.9%	77.7%	78%
Recall	97.8%	78.3%	79.2%
F_1 measure	97.8%	77.9%	77.8%

A. The learning curves

One of the challenges that are associated with machine-learning classifiers is bias. Bias in this study is considered as model's inability to identify correctly a true association between the target (churn) and the attributes. A high biased model can be a result of underfitting. In the same vein, high variance in data sets is a result of overfitting. An accurate model is the one that has low bias and low variability. To analyze bias and variability of the models, we made use of the learning curves. The learning curve shows the performance measure of the test set on a varied number of training samples.

1) Logistic Regression's learning curve

Fig. 4 indicates the learning curve for the logistic model. This learning curve shows a relatively high variability until about 5000 instances, shown by the larger shade around the testing score. The testing score is much lower until 5000 instances when it begins to approach convergence. The training score and the testing score converge at about 0.83 F1 score and at about 5500 instances. What this could mean is that, after 5500 instances, the model is no longer benefiting much, in other words it has learned enough to be in a position to predict from new data.

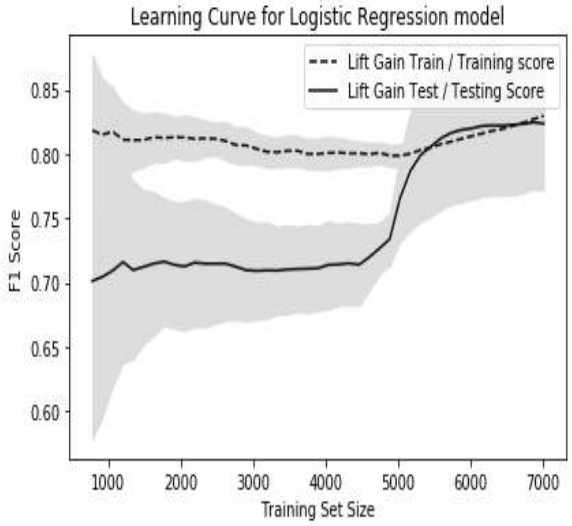

Fig. 4. *The Logistic Regression's learning curve*

2) Decision Tree's learning curve

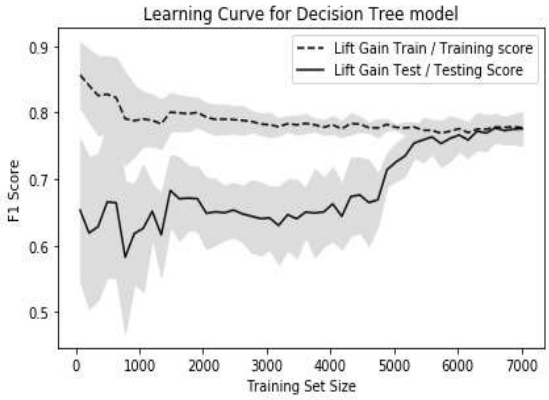

Fig. 5. *Learning curve (Decision Tree)*

Fig. 5 shows the learning curve for the decision tree model. From the graph we can see that the shaded area around the testing score is relatively large until about 4800 instances. This shows that there is relatively high-test variability until about 4800 instances. Between, approximately 1500 and approximately 4800 instances, the test score is very much lower. This implies that the model still needs more training data to learn. After about 4800 instances the testing score then shows an increasing score again until convergence at about 6400 instances. This convergence is an indication that model no longer benefits from adding more train data.

3) Random Forest's learning curve

The random forest's learning curve is presented in Fig. 6. This graph shows high test variability up to around 1500 instances. High variance is also confirmed by a huge gap between the testing and the training curves. It can also be noted that the testing score's gain remains increasing as the score approaches convergence. Interestingly, there is no convergence until probably around 15000 instances which falls outside the graph. This is obviously consequential of the training curve's near constant progression as more data is added. This would imply that the model still has a potential to gain if given more training data is added after 7000 instances.

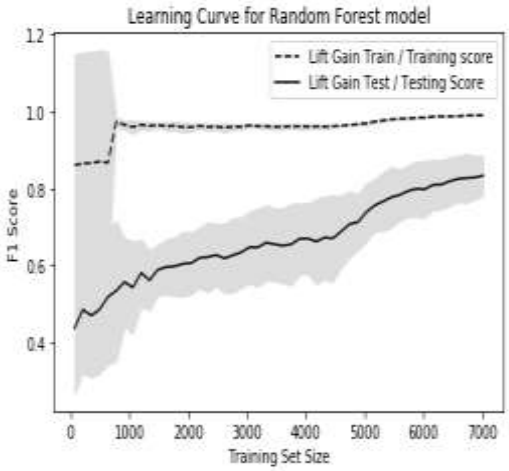

Fig. 6. *The learning curve (Random Forest)*

B. The lift charts

The lift chart is a good measure that also aids in identifying a good prediction model. The idea behind a lift chart is basically to show if the prediction rate of the target attribute is free or independent from random occurrence.

1) Logistic regression's lift chart

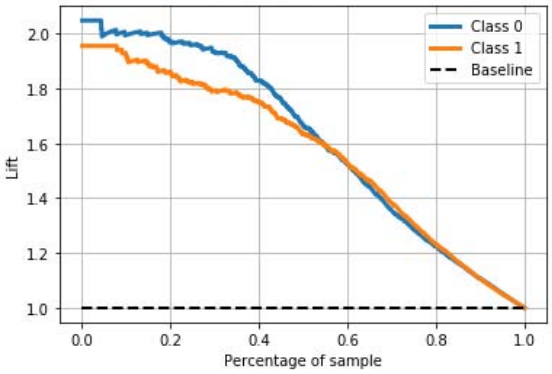

Fig. 7. *The lift chart for the logistic regression model*

Fig 7. indicates the lift chart for the logistic regression model. We can see from the graph that class 1 curve which represents the churners, has a maximum lift of about 2.1. Furthermore, the graph shows that there is an ideal lift for class 1 (churners) that stretches to about 0.05 percentage of the sample. The ideal lift for class 0 (non-churners) stretches to about 0.01 percentage of the sample.

2) Decision Tree's lift chart

From Fig. 8 we can see that both curves drop sharply and constantly from the lift of 2.0 to about 1.4 for the churners class and for about 1.38 for the class of non-churners. Both curves then suddenly increase their lift scores before they start dropping again. The maximum lift for the decision tree is 2.0 as presented in Fig. 8.

3) Random Forest's lift chart

The random forest's lift chart is indicated in Fig. 9. This graph shows a maximum lift point of approximately 2.5. The graph further shows that the graph has an ideal lift, which is the part from the start of the curve where the curve remains constant. This indicates the increase of the population (about 0.02 percentage sample) with a maximum lift, this is the churners population.

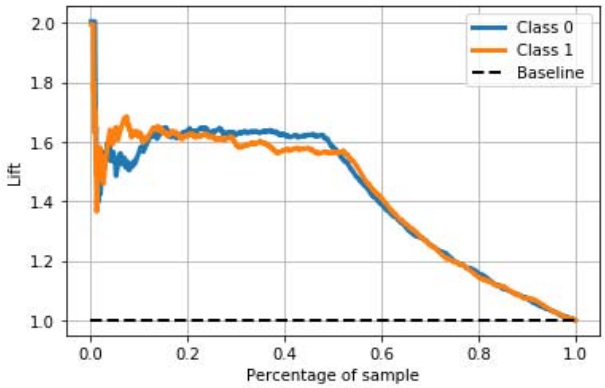

Fig 8. *The Lift chart (Decision Tree)*

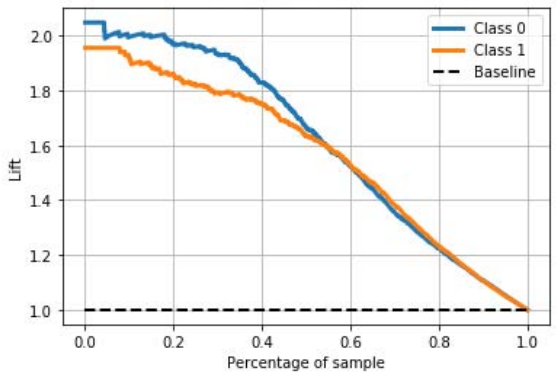

Fig. 9. *Lift chart (Random Forest)*

C. The Precision-Recall curves

Precision-Recall (PR) curve serves the same purpose served by the ROC curve; it measures the success of prediction. However, the ROC curve works well only when the dataset does not have data imbalance. Since this study's dataset is imbalanced, it was more appropriate to use the PR metric.

1) Logistic Regression's PR curve

Fig. 10 presents the logistic regression-based model's PR curve. It can be noted that the model's average precision (AP) is 0.91. This shows a very high level of prediction success of the model.

2) Decision Tree's PR curve

Fig. 11 presents the PR curve for Decision Tree is shown. The graph indicates that the model's area under curve of the precision recall (AUPR) is 0.78. This shows a very high level of prediction success of the model. It is worth mentioning that the AUPR and the AP (shown in the PR curve of the logistic regression) are equivalent variables and they have the same meaning when interpreted.

3) Random Forest's PR curve

Fig. 12 presents the random forest's PR curve. It can be noted that the model's area under curve of the PR (AUPR) is 0.8. This shows a good success of prediction of the model.

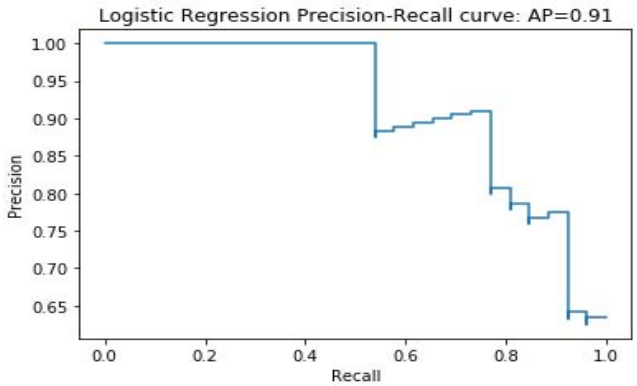

Fig. 10. *The PR curve for the Logistic Regression model*

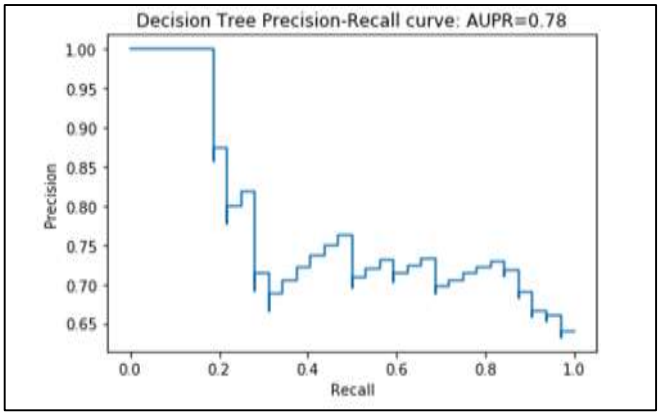

Fig. 11. *The PR curve for the Decision Tree model*

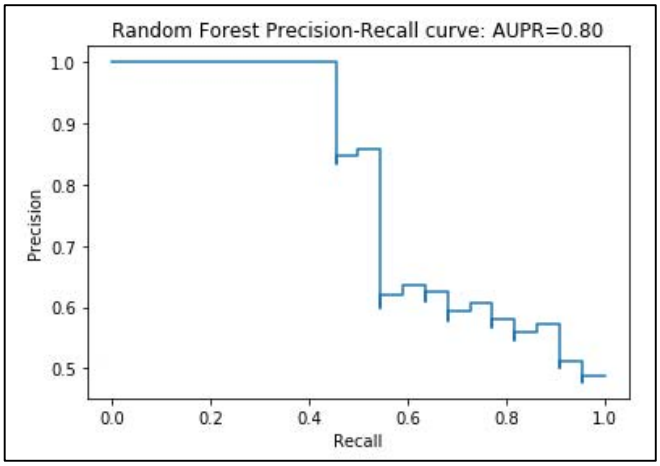

Fig. 12. *The PR curve for the Random Forest model*

D. Results from testing the Logistic Regression model

After training and running the models on the study's dataset, the Logistic Regression model was selected for prototype deployment. The model was selected as it outperformed the other two models; the Random Forest and the Decision Tree. To confirm that the Logistic Regression model can really be considered to be suitable for real time customer churn prediction, we did a prototype deployment of the model to test its performance on new customer records continuously being supplied. The results obtained from testing the proposed model application with new subscriber records are presented in this Section. Fig. 13 shows real time customer churn scoring of the Logistic Regression model after a prototype deployment. The graph reveals that the model was able to correctly predicting the target variable, the likely future churners. This is shown by high churn probability score of 75.26%. This shows that the Logistic Regression-based model can be of great use in the telecommunication industry for business planning, especially when decisions have to be taken based on real time data.

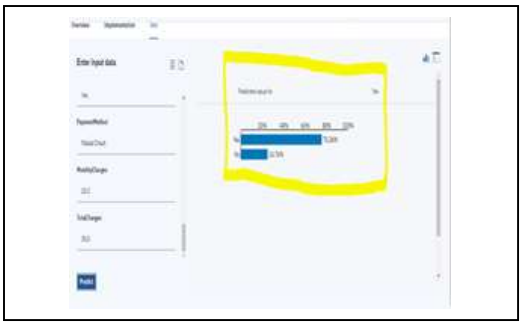

Fig. 13. *Instant (real time) customer churn scoring.*

E. Feature Importance

It is essential for several companies to use various techniques to identify customer attrition (churn) for retaining loyal customers, therefore maximizing revenue. Domain knowledge is required to determine the causes of customer churn. This is implemented by identifying features that greatly contribute to subscribers intentionally leaving the company. Fig. 14 shows how various attributes in the Telco dataset influenced customer churn. Monthly charges, total charges, tenure and weights were the most significant attributes in predicting subscriber churn. A subscriber with high monthly and total charges will not leave the company and such a subscriber is loyal. It was noted that a subscriber with less total and monthly charges would likely churn. Tenure was also influential in predicting customer churn. Customers with a long tenure for example 5 years were not likely to churn. It is therefore safe to conclude that monthly charges and tenure are the best attributes to use for customer churn prediction.

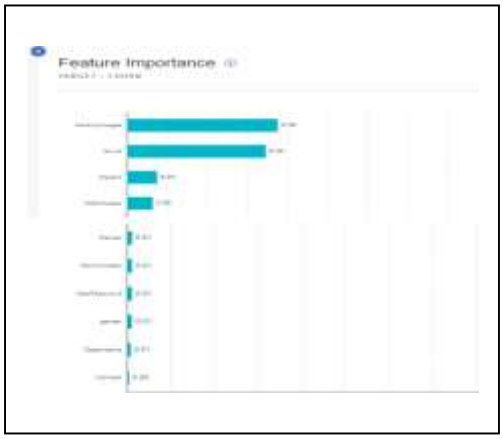

Fig. 14. *The feature importance*

IV. DISCUSSION

This section aims to evaluate the models' performance to determine the best performing model, which would subsequently be deployed as a prototype. The section further aimed at identifying the features that can be indicative of customer churn. The study used the confusion matrix to unpack a number of performance measures. Besides the confusion matrix metric, other measures that were used include the learning curve, lift curve as well as the precision-recall (PR) curve. TABLE 2. indicates that the Logistic Regression-based model outperformed the other two models. Its accuracy was better than that of the second best; the Random Forest, by a whopping 18.6%. The results further showed that all the models had some degree of misclassification, however the misclassification rate of the Logistic Regression was very minimal (2.2%) as differentiated to the Random Forest and the Decision Tree, which had misclassification rates of 20.8% and 21.7% respectively. One interesting thing to note was that both Random Forest and the Decision Tree had good accuracy rates of 78.3% and 79.2% respectively, although still not better than that of the Logistic Regression. Despite the two having good accuracy rates, they had the highest rates of misclassification of class events.

V. CONCLUSION

It was noted from this study that accuracy is not a dependable measure for determining model performance when analyzing imbalanced / unbalanced datasets. Furthermore, the Logistic Regression was shown to have a better sensitivity rate, making this model the best of the three in correctly predicting the churn event when it is actually the churn event, meaning that it rarely misclassified the minority class. Bias is undesirable in any prediction model, concurrently, it is inescapable. Therefore, the best model would exhibit the smallest probability of bias towards the majority class. On that, the Logistic Regression model was shown to have a better sensitivity rate (97.8%) than the other two. Higher sensitivity rate implies less bias. However, the Logistic Regression was shown to suffer from variability, which could indicate overfitting. The issue of overfitting was investigated and the study could not find anything else than variability to support overfitting. In fact, high accuracy rate of the model showed that it was not overfitting, hence we reluctantly concluded that high variability might have been a result of a weak continuous covariate to a model. Investigate if accuracy and variability of a model indeed cannot happen simultaneously is part of the future work. Literature suggests that these two are parallel, but we experienced a case whereby the learning curve of the Logistic regression showed some considerable variability yet the model was more accurate than the other two models, the Decision Tree and the Random Forest-based models. Furthermore, this study will be expanded in future to use a bigger dataset covering a longer time span with the focus to study the issue of churn indicative features in depth.

ACKNOWLEDGMENT

This research was done at the University of Fort Hare, Department of Computer Science sponsored by the Telkom Centre of Excellence and the NRF (National Research Foundation). The ideas, deduction and discoveries articulated in this research belong to the researcher and takes full responsibility.

REFERENCES

[1] ICASA report, 2020. "The State of the ICT Sector Report in South Africa", © ICASA's report on the state of ICT sector in SA, accessed 29 May 2020, <https://www.icasa.org.za/uploads/files/State-of-the-ICT-Sector-Report-March-2020.pdf >.

[2] Kaur, M. and Mahajan, P., 2015. Churn prediction in telecom industry using R. *Int. J. Eng. Tech. Res.(IJETR)*, *3*(5).J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., Dai, W., Yang, Q. and Zeng, J., 2015, May. Telco churn prediction with big data.

In *Proceedings of the 2015 ACM SIGMOD international conference on management of data* (pp. 607-618). ACM.

[4] Ahmad, A.K., Jafar, A. and Aljoumaa, K., 2019. Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), p.28

[5] He, Y., He, Z. and Zhang, D., 2009, August. A study on prediction of customer churn in fixed communication network based on data mining. In 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery (Vol. 1, pp. 92-94). IEEE.

[6] Dahiya, K. and Bhatia, S., 2015, September. Customer churn analysis in telecom industry. In 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions) (pp. 1-6). IEEE.

[7] Brândușoiu, I., Toderean, G. and Beleiu, H., 2016, June. Methods for churn prediction in the pre-paid mobile telecommunications industry. In 2016 International conference on communications (COMM) (pp. 97-100). IEEE.

[8] Cloud.ibm.com. n.d. IBM Cloud, accessed 8 October 2019, https://cloud.ibm.com/login

[9] Piatetsky, G., 2014. CRISP-DM, still the top methodology for analytics, data mining, or data science projects. KDD News.

[10] Azevedo, A.I.R.L. and Santos, M.F., 2008. KDD, SEMMA and CRISP-DM: a parallel overview. IADS-DM.

[11] Kaggle.com., 2018. WA_Fn-UseC_-Telco-Customer-Churn, accessed 10 May 2019, <https://www.kaggle.com/palashfendarkar/wa-fnusec-telcocustomerchurn>.

[12] Blalock Jr, H.M., 1963. Correlated independent variables: The problem of multicollinearity. Social Forces, 42(2), pp.233-237.

[13] Kelleher, J.D., Mac Namee, B. and D'arcy, A., 2015. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press.

[14] Lakshminarayan, K., Harp, S.A. and Samad, T., 1999. Imputation of missing data in industrial databases. Applied intelligence, 11(3), pp.259-275.

[15] Yemulwar, S., 2019. Feature Selection Techniques, accessed 14 June 2020, <https://medium.com/analytics-vidhya/feature-selection-techniques-2614b3b7efcd>.

[16] Maheshwari, S., Jain, R.C. and Jadon, R.S., 2017. A Review on Class Imbalance Problem: Analysis and Potential Solutions. *International Journal of Computer Science Issues (IJCSI)*, *14*(6), pp.43-51.

[17] Chawla, N.V., 2009. Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook (pp. 875-886). Springer, Boston, MA.

[18] Lohr, S.L., 2009. *Sampling: design and analysis*. Nelson Education.

[19] Hashmi, N., Butt, N.A. and Iqbal, M., 2013. Customer churn prediction in telecommunication a decade review and classification. International Journal of Computer Science Issues (IJCSI), 10(5), p.271.

[20] Ibm.com. n.d. IBM Knowledge Center, accessed 14 November 2019, <https://www.ibm.com/support/knowledgecenter/en/SS6NHC/com.ibm.swg.im.dashdb.doc/overview.html>.

[21] Miller, J.D., 2019. Hands-On Machine Learning with IBM Watson: Leverage IBM Watson to implement machine learning techniques and algorithms using Python. Packt Publishing Ltd.