International Conference on "Interdisciplinary Research in Technology & Management" (IRTM 2021)

Conference Proceedings

Edited by

Satyajit Chakrabarti
Director, IEM Kolkata, India

Rintu Nath
Scientist – F, Vigyan Prasar, Department of Science and Technology, Govt. of India

Pradipta Kumar Banerji
Dean (Management Studies), Institute of Engineering and Management, Kolkata

Sujit Datta
Institute of Engineering & Management, Kolkata, India

Sanghamitra Poddar
Institute of Engineering & Management, Kolkata, India

Malay Gangopadhyaya
Institute of Engineering & Management, Kolkata, India

First edition published 2022
by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

and by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

CRC Press is an imprint of Informa UK Limited

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalogue entry has been requested.

ISBN: 978-1-003-20224-0 (ebk)

DOI: 10.1201/9781003202240

Chapter 98

Predicting Telecommunication Customer Churn using Machine Learning Techniques

Daisy Reneilwe Chabumba
School of Computer Science and Applied Mathematics
The University of the Witwatersrand, Johannesburg, South Africa

Ashwini Jadhav
Science Teaching and Learning Unit Faculty of Science
The University of the Witwatersrand, Johannesburg, South Africa

Ritesh Ajoodha
School of Computer Science and Applied Mathematics
The University of the Witwatersrand, Johannesburg, South Africa

Abstract—Customer churn is a major problem and one of the most important concerns for large businesses. Due to the direct effect on the profits of the companies, especially in the telecommunication area, companies are looking for ways to develop means to predict potential customers which may churn. Therefore, looking for factors that increase customer churn is important to take necessary actions to reduce this retention. The main contribution of our work is to develop a churn prediction model which assists telecommunication businesses to predict customers who are most to leave after a certain period of time. The model developed in this work uses machine learning techniques on big data platform and builds a new way of feature 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 84%. The model was prepared and tested through python environment by working on a large dataset that was found at www.kaggle.com. The model experimented four algorithms: Logistic Regression, Random Forest, Support Vector Machines and Extreme Gradient Boosting "XGBOOST". However, the best results were obtained by applying random forest algorithm at a 80% accuracy.

I. INTRODUCTION

Customer churn is the percentages of customers that has stopped using a company's product during a certain period of time. It is calculated by dividing the number of customers a business loses at the end of the period by the number of customers a business had in the beginning of the time frame. Customer churn is one of the most important factors for a growing business to evaluate. This is the number that can give a company direction and insights [1].

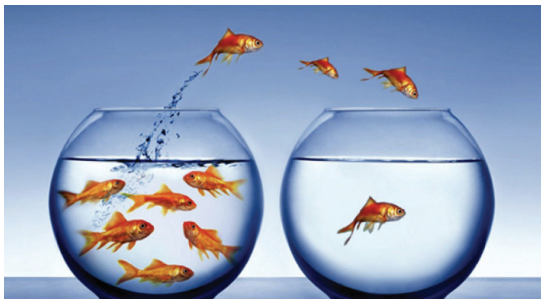


Fig. 1. Visual interpretation of the customer churn adopted from google.

Customer churn has always been a major problem and one of the biggest concerns for large businesses. Due to the direct effect it has on the profits of a company, more especially in a telecommunication area, companies have went out to look for ways to develop a system that would help to predict which customers were most likely to leave. Since the telecommunication field is a very competitive area, many competitors within the industries, have resorted in finding ways to keep clients. Studies have shown that it is more costly to get new clients that to keep the existing ones [1].

We proposed a predictive approached where a large telecommunication data is used to develop models which are capable of predicting, classifying and explaining the customer churn problem. We used telecommunication customer churn data from www.kaggle.com for this particular study. Machine learning techniques such as support vector machines, random forests and logistic regression and XGBoost classifier were used to give better results due to their strong classification nature. We used confusion matrix as well as the area under the curve (AUC) to evaluate our results.

In this study, we would like to find out which customers are most likely to churn? Can we predict which factors affect the customer churn? Can we predict which customers should get automated renewals before their contracts expire? What impact does customer churn have on the business? Can we identify the features that has the greatest impact on the customer churn rate?

There have been several attempts to predict customer churn using large company data. Hung (2006), He (2009), Huang (2015) and Brandusoiu (2016) gave the best results with the highest accuracy showing that the four predictive models which are support vector machines, random forests and logistic regression and XGBoost classifier are the best to use in order to get the best results. Hung (2006) used data from a wireless company, meanwhile He (2009) used data of 5.23 Million customers. Huang (2015) used data from the operations support and business support department of a telecommunication company. Brandusoiu (2016) used data-sets for call details of 3333 customers from a company.

In this study used the Cross Industry Standard Process for Data Mining which follows the following six-step process [17]:

- Understanding the domain and developing the goal for the research;
- Identifying, assessing and understanding the relevant data sources;
- Preprocessing, cleaning, and transforming relevant data;
- Developing methods using comparable analytic techniques;
- Evaluating and assessing the validity and the utility of the models against each other; and
- Deploying the models for use in decision making process.

This favoured methodology gives a systematic and structured way of conducting data mining studies and thus enlarge the possibility of acquiring accurate and authentic results.

This study will make contributions through the body of literature by:

- Identifying which customers are most likely to churn.
- Knowing which factors to look out for with regards to customer churn.
- Knowing which customers to give immediate attention to to avoid customer churn.
- Identifying which features has a great impact on the customer churn rate.

The rest of the report is arranged in this sequence. The next section is the related work which will be followed by the methodology. Then in section 5, we have discussion and results. The last section will present the conclusion and future work.

II. RELATED WORK

In this section, we see the work that the other authors have done which may give light to our study. Understanding predictor variables that affects customer churn is crucial. We will see the data that was used, the models which the authors opted for, the accuracy of those models and how the models were evaluated.

A. Features

Some of these authors have similarities in their studies. He (2009) and Zhang (2012) used a large and feature rich for their study. He (2009) considered 5.23 million customers of a large Chinese telecommunication company. Zhang (2012) used experimental data from a leading service provider and compromise more than 1 million customers. Huang (2012) considered data of 827 124 customers that were randomly selected from a real world database provided by the telecommunications in Ireland. Buckinx (2005) considered retailing data-sets of 158 884 customers.

Sharma (2011) and Brandusoiu (2016) used voice calls data. Sharma (2011) considered churn data-set from the UCI Repository of ML databases at the University of California Irvine. The churn data-set deals with cellular service providers and the data pertinent to voice calls they make. Brandusoiu (2016) considered the data-set for call details of 3333 customers with 21 features, and a dependent churn parameter with values: yes/no. Huang (2015) used departmental data. Huang (2015) considered the operation support department and business support department in China's largest telecommunication company.The rest of the other authors used data from institutional data bases. Hung (2006) considered data from a wireless telecommunication company and Coussement (2008) considered the subscriber data-set of an institution.

B. Models

There is a limited number of models that can be used in Machine Learning. Some of these models give the best results while some give poor results. We see that a lot of authors used the most popular models. Artificial neural networks were used by almost all the authors. Buckinx (2005), Coussement (2008), Huang (2012) and Huang (2015) used the random forests. Buckinx (2005) and Coussement (2008) also used the logistic regression.

Huang (2012) considered a number of predictive models including, Bayes Naves, decision trees, linear classifiers as well as support vector machines. Hung (2006), He (2009), Sharma (2011), Zhang (2012) and Brandusoiu (2016) used the artificial neural networks. Huang (2006) also used decision trees and K means clustering meanwhile Zhang (2012) considered logistic regression and decision trees. Brandusoiu (2016) also used the support vector machines and the Bayes network. We see that the artificial neural networks gave the best results.

C. Accuracy

The models that were used gave the results, from poor to rich results. We see that Hung (2006), He (2009), Sharma (2011), Huang (2015), and Brandusoiu (2016) gave the best results with the highest accuracy showing that the two predictive methods which are the artificial neural network random forests are the best to use in order to get best results. We see that the random forests and artificial neural networks have a tough competition. Sometimes the support vector machines give the highest accuracy followed by the decision trees and visa versa. We have also seen that the logistic regression are greatly used. The Bayes networks, clustering and Bayes Naive were overpowered by the other models.

The highest accuracy in the algorithm is the reason why we choose to work with random forests and artificial networks. Due to the logistic regression being the mostly used classification techniques, which has its roots in traditional statistics, we have decided to used this technique to see what results it will give.

D. Evaluation Functions

Taking the results of the data mining models without any evaluation processes can be risky and lead to wrong decision making [7]. Different authors use different evaluation for the performance of the model. We see that the confusion matrix was one of the most used evaluation tool. Huang (2012) used the confusion matrix, Receiver Operating Characteristic curve (ROC) and the Area under the ROC curve (AUC). Sharma (2011) and He (2009) only considered the confusion matrix. Brandusoiu (2016), Coussement (2008) and Huang (2015) considered AUC to measure the performance of the algorithm. Hung (2006) considered top-decile lift and hit ratio. Buckinx (2005) used the classification accuracy and the area under the curve (AUC). Zhang (2012) used top-decile lift, AUC and hit ratio.

The following table is the summarized version of the related work.

Authors	Data / Feature	Models	Evaluation / Accuracy
Buckinx and Van den Poel (2005).	Considered retailing data-sets of 158 884 customers.	Artificial neural networks, logistic regression and random forests.	Random forests outperformed the other techniques.
Hung, Yen and Wang (2006).	Considered data from a wireless telecommunication company.	Decision trees, artificial neural network and K means clustering.	The artificial neural network the best predictive results at 99% and the decision trees gave the second best at 90%
Coussement and Van den Poel (2008).	They considered the subscriber data-set of an institution.	Support vector machine, random forest and logistic regression.	Random forests out performed the other techniques.
He, He and Zhang (2009).	They considered 5.23 million customers of a large Chinese telecommunication company.	Artificial neural network.	The artificial neural networks gave the predictive results of 91.1%.

Sharma and Panigrahi (2011).	They considered churn data-set from the UCI Repository of ML Databases at the University of California Irvine. The churn data-set deals with cellular service providers and the data pertinent to voice calls they make.	Artificial neural network.	The predicted accuracy for the artificial neural network is 92.35%.
Zhang, Wan, Xi and Zhu (2012).	Experimental data from a leading service provider and com-promise more than 1 million customers.	Logistic regression, artificial neural networks and decision trees.	The results showed Artificial neural networks outper-forms the logistic regression while the logistic regression outper-forms the decision trees.
Huang, Kechadi and Buckley (2012).	They considered data of 827 124 customers that were randomly selected from a real world database provided by the telecommunications in Ireland. Each customer is represented by 738 features.	Bayes Naive, decision trees, random forests, artificial neural networks, logistic regression, linear classifiers, and support vector machines.	The C4.5 decision trees and the support vector machines gave the best predictive results, outperforming the other modelling techniques for churn prediction.
Huang, Zhu, Yuan, Deng, Li and Ni (2015).	They considered the operation support department and business support department in China's largest telecommunication company.	Random forests.	The results showed that random forests gave a 90% overall accuracy.
Brandusoiu, Toderean and Beleiu (2016).	They considered the data-set for call details of 3333 customers with 21 features, and a dependent churn parameter with values: yes/no.	Artificial neural network, support vector machines and Bayes network.	The highest prediction accuracy was 99.70% obtained with Bayes network algorithm, followed by the support vector machines. The artificial neural network came last with 99.10%

III. RESEARCH METHODOLOGY

This study will be using the Cross Industry Standard Process for Data Mining which follows the following six-step process [17]:

- Understanding the domain and developing the goal for the research;
- Identifying, assessing and understanding the relevant data sources;
- Preprocessing, cleaning, and transforming relevant data;
- Developing methods using comparable analytic techniques;
- Evaluating and assessing the validity and the utility of the models against each other; and
- Deploying the models for use in decision making process.

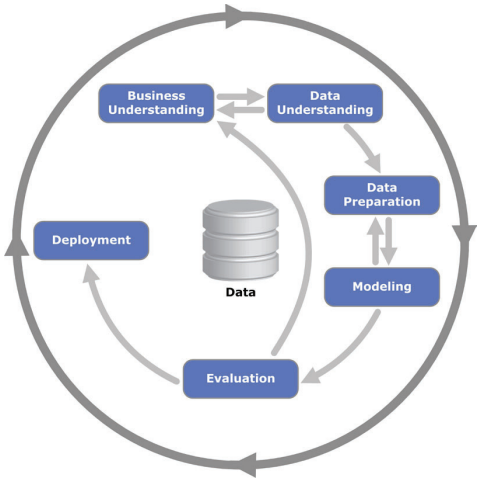
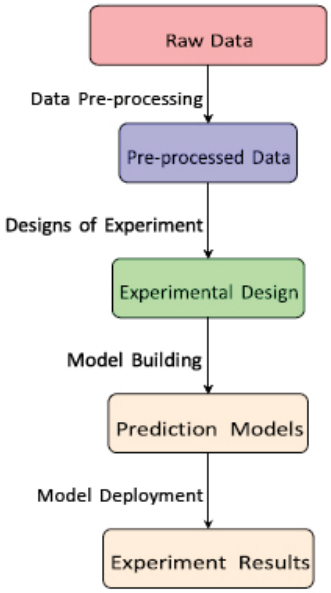


Figure 2: Visual interpretation of the cross industry standard process for data mining adopted from google.

This favoured methodology gives a systematic and structured way of conducting data mining studies and thus enlarge the possibility of acquiring accurate and authentic results. The fact that we start by understanding the domain followed by understanding the data and preparing the data this positions the study to be a prosperous data mining study. In this study we will use the 10-fold cross-validation approach to estimate the performance of the prediction models. We set the k into 10 since studies show that 10 appears to be an optimal number of folds [11].

A. Step-by-step processes for the study

The Data Mining and Cross-Validation process.



We started by getting the raw data from the IBM telecommunication data-set available from www.kaggle.com then processed the data to get pre-processed data. We then applied designs of experiment to get to the experimental design where we had 10-fold cross-validation. We did model testing by applying confusion matrix and the area under the curve (AUC). From the experimental design we had model building using our prediction models.

B. Data description

We used the IBM telecommunication data-set available from www.kaggle.com. The raw data contained 7043 rows (customers) and 21 columns (features) [10]. Each row represents a customer, each column contains customer's attributes described on the column Metadata.

The data-set was divided into the following categories:

- Customers who left within the last month – the column is called Churn.
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
- Customer account information – how long they have been a customer, contract, payment method, paperless billing, monthly charges, and total charges.
- Demographic info about customers – gender, age range, and if they have partners and dependents.

The "churn" column was the target variable. The classification goal was to predict whether the customer will churn.

1) Attributes : Description of variables in the data-set:

- customerID: Customer ID
- gender: Whether the customer is a male or a female
- SeniorCitizen: Whether the customer is a senior citizen or not (1, 0)
- Partner: Whether the customer has a partner or not (Yes, No)
- Dependents: Whether the customer has dependents or not (Yes, No)
- tenure: Number of months the customer has stayed with the company
- PhoneService: Whether the customer has a phone service or not (Yes, No)
- MultipleLines: Whether the customer has multiple lines or not (Yes, No, No phone service)
- InternetService: Customer's internet service provider (DSL, Fiber optic, No)
- OnlineSecurity: Whether the customer has online security or not (Yes, No, No internet service)
- OnlineBackup: Whether the customer has online backup or not (Yes, No, No internet service)
- DeviceProtection: Whether the customer has device protection or not (Yes, No, No internet service)
- TechSupport: Whether the customer has tech support or not (Yes, No, No internet service)
- StreamingTV: Whether the customer has streaming TV or not (Yes, No, No internet service)
- StreamingMovies: Whether the customer has streaming movies or not (Yes, No, No internet service)
- Contract: The contract term of the customer (Month-tomonth, One year, Two year)
- PaperlessBilling: Whether the customer has paperless billing or not (Yes, No)
- PaymentMethod: The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic),
- Credit card (automatic))
- MonthlyCharges: The amount charged to the customer monthly
- TotalCharges: The total amount charged to the customer
- Churn: Whether the customer churned or not (Yes or No)

C. Data Pre-processing

This section will diligently talk about the data pre-processing. Data pre-processing is the step before applying data mining algorithm where the original data is transformed into a suitable shape to be used by a particular mining algorithm. Data pre-processing includes different tasks as data cleaning, feature selection and data transformation. [16].

1) Data Cleaning: Data cleaning is one of the most important pre-processing tasks, which is applied on this data set to remove irrelevant items and missing values.

Firstly, we got rid of columns that are unnecessary. The customerID colomn was removed because it did not add any more value to the model or the analysis of the problem since it is just an ID for the customer. With the one column removed, we were left with 7,043 rows and 20 columns as the new data-set.

We then converted all non-numeric columns or categorical columns to numerical columns. To check if this was done successfully, we then checked the data types of the new data.

```
gender              int64
SeniorCitizen       int64
Partner             int64
Dependents          int64
tenure              int64
PhoneService        int64
MultipleLines       int64
InternetService     int64
OnlineSecurity      int64
OnlineBackup        int64
DeviceProtection    int64
TechSupport         int64
StreamingTV         int64
StreamingMovies     int64
Contract            int64
PaperlessBilling    int64
PaymentMethod       int64
MonthlyCharges    float64
TotalCharges        int64
Churn               int64
```

Figure 3: All numerical data types from the new data set

We then checked if the data is good to use and if all of the data types are numerical values.

We split the data into training and testing data sets. 80% of the original cleaned data was deployed for the training data the model and 20% was deployed for the testing data the model.

2) Feature Selection: Feature selection is a fundamental task in data pre-processing area. This is also an important task done under data pre-processing. The objective of feature selection process is to select an appropriate subset of features which can perfectly describe the input data, reduces the dimensionality of feature space, removes unnecessary and irrelevant data [9].

This process can play an important role in improving the data quality therefore the performance of the learning algorithm. Feature selection methods are categorized into wrapper-based and filter-based methods. Filter method is searching for the minimum set of relevant features while ignoring the rest. It uses variable ranking techniques to rank the features where the highly ranked features are selected and applied to the learning algorithm. Different feature ranking techniques have been proposed for feature evaluations such as information gain and gain ratio [9].

In this study, we applied filter-method using information gain based selection algorithm to evaluate the feature ranks, checking which features are most important to build customer churn model. During feature selection, each feature assigned a rank value according to their influence on data classification. The highly ranked features are selected while others are excluded. Total charge feature got the higher rank, then followed by tenure, monthly charges, contract (month-to-month) and online security features. As we can see the appropriate subset of features consist of five features while other ones are excluded. In summary, the features that are related customer account information have an greater impact on the customer churn.

D. Predictive Models

In this study, four popular classification methods— (support vector machines, random forests classifier, XGBoost classifier and logistic regression) are built and compared to each other making use of their predictive accuracy on the given data samples.

A sizeable number of studies compare predictive methods. Most of these previous studies found machine learning methods (i.e support vector machines and random forests) to be superior to their statistical equivalent (i.e logistic regression) in terms of both being less constrained by assumptions and producing better prediction results. From the literature that was reviewed, Hung (2006), He (2009), Sharma (2011), Huang (2015), and Brandusoiu

(2016) gave the best results with the highest accuracy showing that the two predictive methods which are the artificial neural network and random forests are the best to use in order to get best results. We used logistic regression, random forests classifiers and the support vector machines as informed by the literature. We decided to add XGBoost classifier because it is an efficient and easy to use algorithm which delivers high performance and accuracy as compared to other algorithms.

1) Logistic Regression: Logistic regression is the mostly used classification techniques, which has its roots in traditional statistics. The concept of the logistic regression is to examine the linear relation between the dependent variable and the independent variable. The dependent variable may be binomial or multi-nomial [19]. Logistic regression is a predictive analysis.

Logistic regression is used to describe data. Some of the advantages of using this method include: logistic regression is one of the simplest machine learning algorithms and is very much easy to use and implement yet provides great training efficiency in some cases. This algorithm proves to be very efficient when the data-set has features that can be separated linearly. The algorithm allows models to easily update to reflect new data, unlike the decision trees or support vector machines. The logistic regression algorithm can easily be extended to multi-class classification using a soft max classifier [14].

Some of the disadvantages of using logistic regression includes: it is difficult to see complex relationships using logistic regression. Only important features should be used to build a model otherwise the probabilistic predictions made by the model may be incorrect and this may end in one getting incorrect predictive values. The algorithm is sensitive to outliers hence the presence of data values that deviate from the expected range in the data-set may lead to incorrect results. logistic regression requires a large data-set, it does not really work well with small data-sets [14].

2) Random Forests Classifier: Random forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees [6]. Random forests correct for decision trees' habit of over-fitting to their training set [8] .

Some of the advantages of using this method include: random forest can automatically take care of missing values. The algorithm is powerful in handling outliers. A random forest algorithm is very stable. Non linear parameters don't affect how the a random forest performs unlike any other algorithms. Random forests works well with both categorical and continuous data. The algorithm also reduces over-fitting and the variance which therefore improves the accuracy of the model [12]. Some of the disadvantages of using this method include: random forests creates a lot of trees and combines their outputs. This algorithm requires much more computational power and resources. Random forest require a lot of time to train and makes decision based on the majority [12].

3) Support Vector Machines: Support Vector Machines (SVM) are a technique which is good for both classification and regression problems but mainly preferred for classification problems. Inputs are evaluated according to where they sit in the hyper-plane in the feature space and projected to a (0,1) interval which can be interpreted as possibilities of class membership. The projection is a major component of the algorithm. SVM's represents the best line as the one which has the maximum margin. The maximum margin classification has an additional benefit. Only the closest data points to the line have to be remembered to be able to specify the model and be able to classify all points. These data points are called support vectors [5].

The advantages of support vector machines are: It is effective in high dimensional spaces. Still effective in cases where number of dimensions is greater than the number of samples. Uses a subset of training points in the decision function (called *support vectors*), so it is also memory efficient. Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels. The disadvantages of support vector machines include: If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial. SVM's do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

4) XGBoost Classifier: XGBoost is an optimized disbursed gradient boosting library designed to be quite efficient, bendy and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting which helps solve many data science problems very fast and in an accurate manner.

XGBoost is a good algorithm. It works well with most types of data. The algorithm works well on small data, data with subgroups, big data, and complicated data. It does not

work so well on sparse data. However, it tends to do better than most supervised learning algorithms on those types of data problems. This algorithm is efficient in handling missing data. XGBoost is an easy to use algorithm which deliverers high performance and accuracy compared to other algorithms. The biggest limitation is probably it's black box nature. It does not give effect sizes. One would need to derive and program that part yourself.

E. Evaluation

When using machine learning algorithms we need to test to see if the identified models work. We propose to use the confusion matrix also known as an error matrix [18], which is a specific table layout that permits visualization of the performance of an algorithm, typically a supervised learning one. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa) [15].

IV. RESULTS AND DISCUSSION

A. Interpretation of Results

In this section we show graphics of the data that was used. We see the results directly from the data and how it is affected from customer churn. About 5,174 customers did not churn and 1,869 customers left the company. We then showed this count visually.

The below visuals gave us an idea of which feature has an effect on customer churn and which one does not. They also show that customers who have not signed up for certain things are most likely to churn. Those that did not have an effect on the customer churn included: gender, dependents, partner, phone services, multiple lines, streaming TV, payment method and senior citizen.

Meanwhile the customers that did not sign up for the following: online security, online backup, device protection, tech support and paperless billing are most likely to churn. Customers who signed up for fibre optics are most likely to churn, about 50% have churned. Customers without internet service have a low rate of churn.

Contracts have a great impact on customer churn. The month-to-month contract had the most customer churn. Not churned customers have a longer average tenure of about 20 months than churned customers. Churned customers paid over 20% higher on average monthly fees than not churned. Churned customers paid more than not charged.

The following graphs are derived directly from the data.

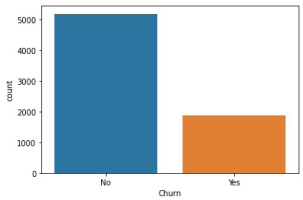


Fig. 4. Visual count of customer churn

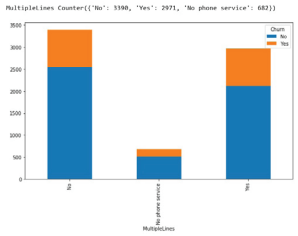


Fig. 5. Plot for the Multiple Lines

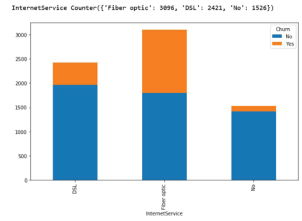


Fig. 6. Plot for the Internet Service

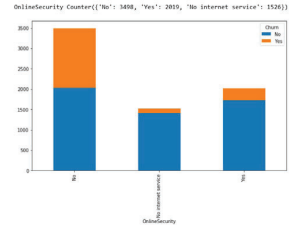


Fig. 7. Plot for the Online Security

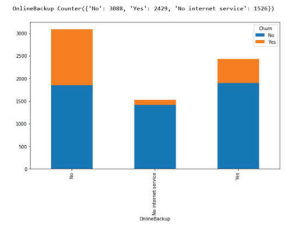


Fig. 8. Plot for the Online Backup

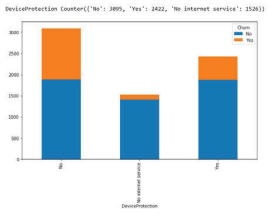


Fig. 9. Plot for the Device Protection

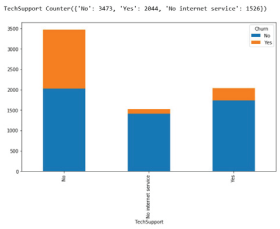


Fig. 10. Plot for Tech Support

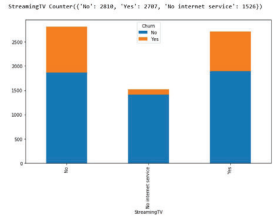


Fig. 11. Plot for Streaming TV

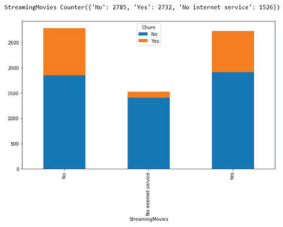


Fig. 12. Plot for Streaming Movies

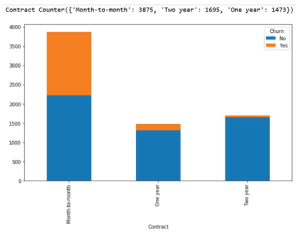


Fig. 13. Plot for Contracts

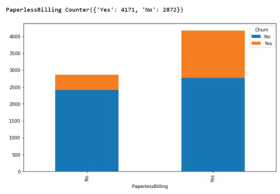


Fig. 14. Plot for Paperless Billing

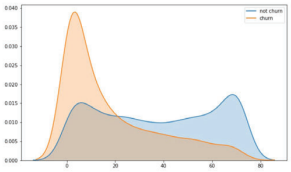


Fig. 15. Plot for Tenure

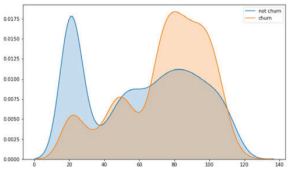


Fig. 16. Plot for Monthly Charges

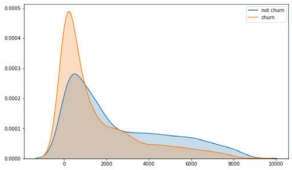


Fig. 17. Plot for Total Charges

B. Evaluation Measures

In our study, we use four common different measures for the evaluation of the classification quality: Accuracy, Precision, Recall and F-Measure [15], [3]. We also used the AUC-ROC curve which is a performance measurement for classification problem. ROC is a probability curve and AUC represents degree of disconnectedness. It tells how much model is capable of distinguishing between classes.Accuracy is the proportion of the total number of predictions where correctly calculated.

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

Precision is the ratio of the correctly classified cases to the total number of misclassified cases and correctly classified cases.

$$Precision = \frac{TP}{TP + FP}$$

Recall is the ratio of correctly classified cases to the total number of unclassified cases and correctly classified cases.

$$Recall = \frac{TP}{TP + FN}$$

In addition, we used the F-measure to combine the recall and precision which is considered a good indicator of the relationship between them [3].

$$F - Measure = 2\frac{Precisionc * Recallc}{Precisionc + Recallc}$$

C. Evaluation Results

There are many features directly or indirectly a_ecting the e_ectiveness of the customer churn model. In this section, we evaluated the impact of customer churn using di_erent classification techniques such as (logistic regression, random forests, support vector machines and XGBoost classifiers).

1.Logistic Regression

Accuracy:	73.80%
Precision Score:	49.94%
Recall Score:	74.33%
F1 score:	59.91%

Confusion Matrix: $\begin{bmatrix} 1146 & 400 \\ 142 & 419 \end{bmatrix}$

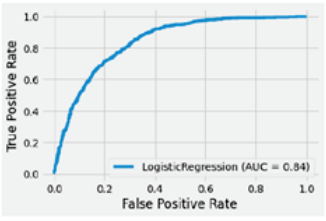


Figure 18: AUC for Logistic Regression Churn Customer.

2.Random Forests Classifier

Accuracy:	80.17%
Precision Score:	65.71%
Recall Score:	49.73%
F1 score:	56.81%

Confusion Matrix: $\begin{bmatrix} 1413 & 193 \\ 277 & 279 \end{bmatrix}$

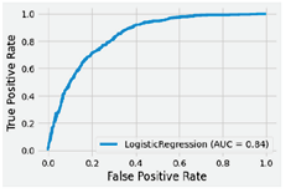
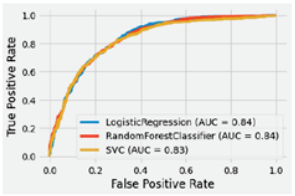


Figure 19: AUC for Random Forest Churn Customer.

3.Support Vector Machine

Accuracy:	79.82%
Precision Score:	67.89%
Recall Score:	45.11%
F1 score:	53.97%

Confusion Matrix: $\begin{bmatrix} 1436 & 130 \\ 308 & 288 \end{bmatrix}$



4.XGBoost Classifier

Accuracy:	78.98%
Precision Score:	61.83%
Recall Score:	50.54%
F1 score:	55.63%

Confusion Matrix: $\begin{bmatrix} 1383 & 178 \\ 275 & 289 \end{bmatrix}$

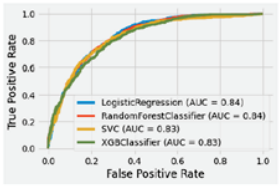


Figure 21: AUC For LR, RFC, SVC and XGBoost.

The table below gives an overview of how each model performed.

	Logistic Regression	Random		
Forests	Support			
Vector				
Machine	XGBoost			
Accuracy	73.80%	80.17%	79.82%	78.98%
Precision	49.94%	65.71%	67.89%	61.83%
Recall	74.33%	49.73%	45.11%	50.54%
f1-score	59.91%	56.81%	53.97%	55.63%
AUC	0.84	0.84	0.83	0.83

Out of all the algorithm, the random forests gave out the more accurate results with an accuracy of 80.17%. This proves that for a classification problem, random forests is the best model to use. Seeing that the random forests gave the highest accuracy, this agrees with the literature that was reviewed.

V. CONCLUSION

Customer churn is a major problem and one of the most important concerns for large businesses. Due to the direct effect on the profits of the companies, especially in the telecommunication area, companies are looking for ways to develop means to predict potential customers which may churn. Therefore, looking for factors that increase customer churn is important to take necessary actions to reduce this retention.

In this study, we proposed a predictive approached where a large telecommunication data is used to develop models which are capable of predicting, classifying and explaining the customer churn problem. We used telecommunication customer churn data from www.kaggle.com for this particular study. Machine Learning techniques such as support vector machines, random forests and logistic regression and XGBoost classifier were used.

In addition, we applied ensemble methods to improve the performance of these classifiers. The obtained results reveal which features to look out for when dealing with customer churn. The results gave the accuracy of 80.17% for random forests, 79.82% for support vector machines, 78.98% for XGBoost and 73.80% for the logistic regression with the AUC of 84%. As expected, the random forest outperformed the other predictive models.

The current methodology of the churn prediction can be tested for other sectors/fields like banking, airline, insurance through comparison for prediction accuracy. This customer churn model can be tested on a very large data, larger than the current.

ACKNOWLEDGMENT

This work is based on the research supported in part by the National Research foundation of South Africa (Grant number: 121835).

REFERENCES

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

[2] Asif, R., Merceron, A., Ali, S.A., and Haider, N.G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers and Education* 113. 177–194.

[3] Chen, T. Y., Kuo, F. C. and Merkel, R. (2004). "On the statistical properties of the f-measure. In Quality Software, 2004. QSIC 2004", Proceedings. Fourth International Conference on. IEEE,pp. 146–153.

[4] Ellis, R. K. (2009). Field Guide to Learning Management System, ASTD Learning Circuits, *archived from the original on 24 August 2014, retrieved 5 July 2012.*

[5] Friedman, J., Hastie, T., and Tibshirani, R. (2009). The elements of statistical learning, volume 1. *Springer series in statistic Springer, Berlin.* ISBN: 978-0-387-84858-7.

[6] Ho, T. K. (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.

[7] Han, J. Kamber, M. (2006). Data Mining: Concepts and Techniques (2nd edition).

[8] Hastie, T., Tibshirani, R., Friedman, J. (2008). The Elements of Statistical Learning (2nd ed.). Springer. ISBN 0-387-95284-5.

[9] Karegowda, A. G., Manjunath, A. S. and Jayaram, M. A. (2010). "Comparative study of attribute selection using gain ratio and correlation based feature selection", International Journal of Information Technology and Knowledge Management, vol. 2, no. 2, (2010), pp. 271277.

[10] https://www.kaggle.com/blastchar/telco-customer-churn

[11] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. The Proceedings of the 14th International Conference on AI (IJCAI), Morgan Kaufmann, San Mateo, CA, 11371145.

[12] Kumar, N. (2019). http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-random.html

[13] Lemmens, A., and Croux, C. (2006). Bagging and boosting classification trees to predict churn, *Journal of Marketing Research* 43(2), 276–286. OpenGenus IQ: Learn Computer Science

[14] Advantages and Disadvantages of Logistic Regression, https://iq.opengenus.org/advantages-and-disadvantagesof-logistic-regression/

[15] Powers, D. M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness Correlation" (PDF). Journal of Machine Learning Technologies. 2 (1): 37–63.

[16] Romero, C., Romero, J. R. and Ventura, S. (2014). "A survey on pre-processing educational data", In Educational Data Mining. Springer International Publishing, pp. 29–64.

[17] Shearer, C. (2000). The CRISP-DM Model: the new blueprint for data mining, Journal of data warehousing 5, 13–22.

[18] Stehman, S. V. (1997). "Selecting and interpreting measures of thematic classification accuracy". Remote Sensing of Environment. 62 (1): 77–89.

[19] Thammasiri, D., Delen, D., Meesad, P., and Kasap, N. (2014). A critical assessment of inbalanced class distribution problem: The case of predicting freshmen student attrition. Volume 41, Issue 2, page 321–330.