March - April 2020
ISSN: 0193-4120 Page No. 2006 - 2012

Analysis of Attrition Prediction Using Machine Learning

[1]Sirisha.Rajanala,[2]K. Swetha
[2]Assistant Professor in Computer Science and Engineering department at
Dr.K.V.SubbaReddy College ofEngineering for Women, Kurnool

Article Info
Volume 83
Page Number: 2006 - 2012
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 18 March 2020*

Abstract
Customer attrition is also known as customer churn, turnoveror defection which means the loss of clients or customers.In Telecommunications sector, they often use customer defection analysis and rates of their defection as a business metric topredict the clients earlier than their churn and try to retain them. This increases the business as cost ofretaining old customers is cost effective than cost involved in getting new customers.
Through this paper we proposed a modelfor predicting the customer churn earlier to win back defending clients using some of the algorithms likeLogistic Regression, Decision trees, Random Forest and some Boosting algorithms such as LightGBM, XGBoost.

1.INTRODUCTION

The mode of communication in the world is done through telecommunications sector. There are many competitors in the telecommunication field, who provide various services to attract new customers. In this win-win competition world, retaining the existing customers by not giving them a chance to get attracted by other service providers is a very big task. So,to increase their business they have toretain their existing customers and do new customer acquisition.To retain an old and existing customer, the churn of the customer should be predicted well in advance. So that the service providers will take necessary actions and try to retain the customer.

Churn of customers is of two types that is voluntary customer churn and involuntary customer churn. In involuntary type of customer churn, customer leaves the service provider unwillingly due to forced situations. But in voluntary type of customer churn, customer willingly wants to leave the service provider due to unsatisfied service or for more exiting offers from other service providers. Generally, service providers concentrate to retain the customers who churn voluntarily. So, to manage the customer churn effectively a company should have an accurate and effective churn prediction model to predict the loss of a client well in advance. If the loss of the client is predicted in advance then the service providers will try to know the reason behind their churn and try to solve their issues and provide more exciting offers to make the customer happy and to retain that customer.

2.PROBLEM STATEMENT

With the ease of communication and a number of telecommunication service providers everyone today has a telecom subscription. Changing of telecom service providers is not an obstacle in the present days because of their attractive and exiting offers. So, this makes the customers of telecom industries easier to churn and telecom service providers are losing many loyal customers.As the telecom users are billion in number even a small fraction of churn leads to high loss of revenue. This is a huge problem being faced by many telecom service providers. In order to address this problem, we came up with a predictive model which identifies the customers churn in advance so that necessary actions are taken by the service

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 2006 - 2012

providers to prevent the loss of valuable customers.

3.LITERATURE SURVEY

A lot of research has been done in the field of CRM (Customer Relationship Management) in various industries for retention of customers and develop strategies to build an efficient model so that specific group of customers can be targeted for retention. Various data mining and statistical techniques have been used for churn prediction of which some famous techniques include Decision trees, Regression Models, Neural Networks, Clustering, Bayesian Models, Support Vector Machine, etc.

One researcher discovered a way to classify clusters and find out predictions through commercial ways in a relational database management system. The researchers have identified two models of predicting customers who have a high possibility to go away. They are 'decision tree' and 'nave Bayes classification' models.

The other researcher, proposed a Neural Network (NN) - based approach to predict customer churn in subscription of cellular wireless services. Furthermore, it was found that when different neural networks topologies were experimented medium- sized NNs performance is essential for the customer churn prediction. some proposed Improved Balanced Random Forests (IBRF) based churn prediction. However, the earlier predictions have some limitations in performance.

4.PROPOSED METHODOLOGY

4.1.DATASET AND FEATURES

The dataset used in this paper consists of call detail record sand is obtained from theUCI repository of Machine Learning. It contains information about the usage of a mobile telecommunication system and has a total number of 7043 customers with 21 variables, of which one variable is the Churn dependent variable with two classes Yes/No.

Customer_ID - Anonymized unique ID, Gender - Categorical (Male/Female), Senior_Citizen-Categorical(0/1), 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, Phone_Service-Whether the customer has a phone service or not (Yes, No), Multiple_Lines-Whether the customer has multiple lines or not (Yes, No, No phone service), Internet_Service-Customer's internet service provider (DSL, Fiber optic, No), Online_Security-Whether the customer has online security or not (Yes, No, No internet service), Online_Backup-Whether the customer has online backup or not (Yes, No, No internet service) ,Device_Protection-Whether the customer has device protection or not (Yes, No, No internet service), Tech_Support-Whether the customer has tech support or not (Yes, No, No internet service), Streaming_TV-Whether the customer has streaming TV or not (Yes, No, No internet service) ,Streaming_Movies-Whether the customer has streaming movies or not (Yes, No, No internet service), Contract-The contract term of the customer (Month-to-month, One year, Two year), Paperless_Billing-Whether the customer has paperless billing or not (Yes, No), Payment_Method-The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)), Monthly_Charges-The amount charged to the customer monthly, Total_Charges-The total amount charged to the customer, Churn-yes or no. The above mentioned are the variables in the dataset.

4.2. DATA MANIPULATION

To the above-mentioned data variables, some manipulations are done like replacing spaces with null values in total charges column,dropping null values,Clean further where there are phrases, senior citizen columnas 0 or 1,Split Churn and not churn,Convert tenureas Tenure_12, Tenure_24,Tenure_48,Tenure_60

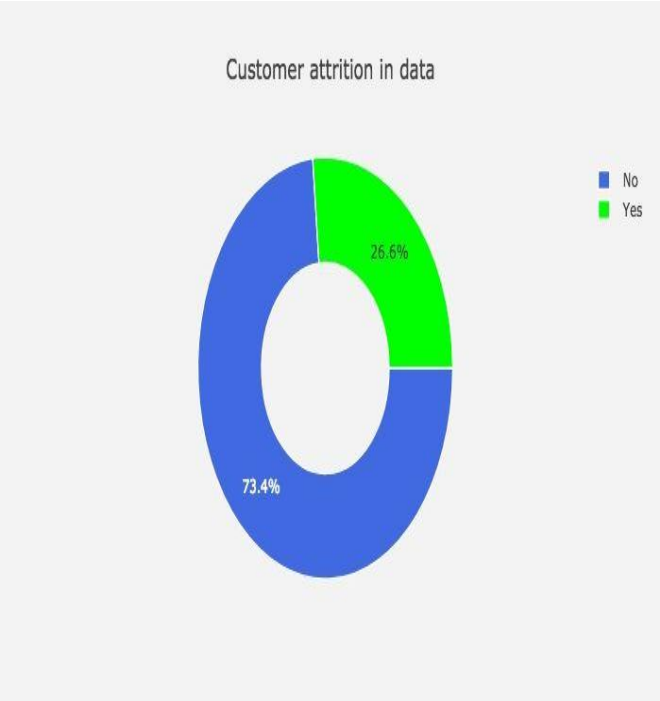
Published by: The Mattingley Publishing Co., Inc.

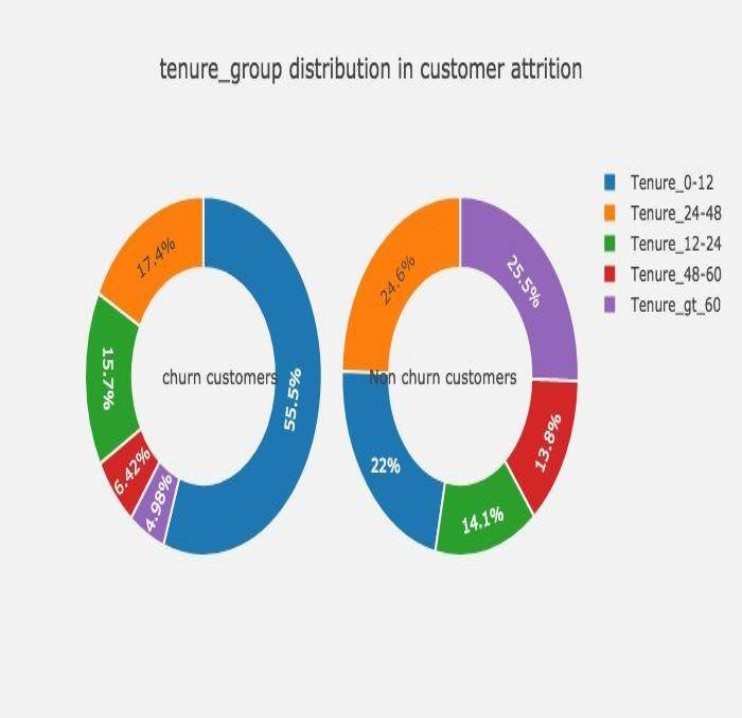
March - April 2020
ISSN: 0193-4120 Page No. 2006 - 2012

4.3. EXPLORATORY DATA ANALYSIS

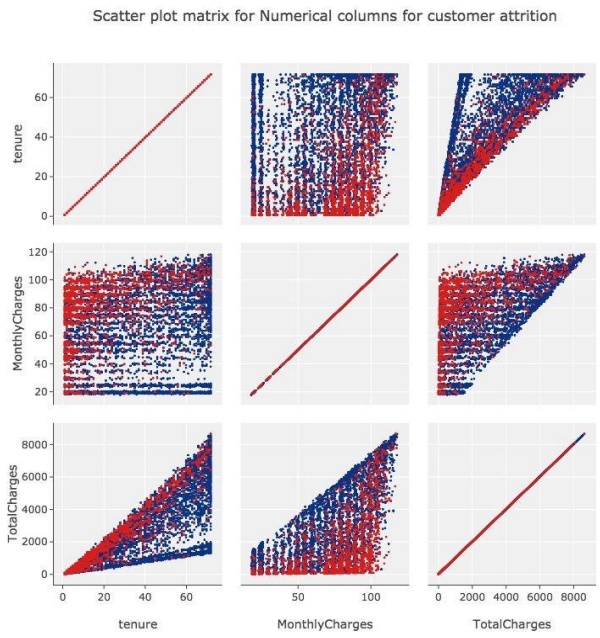
The visualization is done using libraries of python like Pandasand Matplotlib. Data can be understood better by visualizing it. Data visualization is used indata exploration, identifying patterns, detecting outliers and much more. The distribution of the customer attrition in our data as a whole and tenure wise is shown in figure 4.1.

Figure 4.1. customer attrition distribution as a whole and tenure wise





The left plot of fig 4.1 mainly emphasizes that 26.6% of the total customers are churning and 73.4% of customers are still in the service of the company. Imagine the huge billions of profit the company makes if we retain at least 10 to 15 % of the customers who are ready to churn. So, its very important for the companies to predict the churning customers well in advance and to defend them back.The right plot of fig 4.1 emphasizes the proportions of churners and non-churners tenure wise.



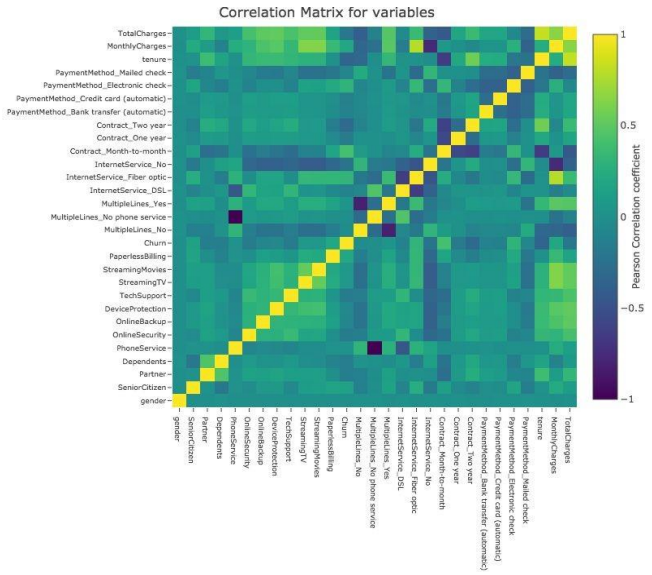
Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 2006 - 2012

Figure 4.2. Scatter plot matrix for continuous variables.

From the fig 4.2. we understand the distribution between continuous variables i.e. how one continuous variable is related to other continuous variables.

Figure 4.3 Pearson Correlation matrix among data attributes.



The features are normalized and then Pearson correlation is calculated. Pearson correlation matrix says about the covariance and standard deviation between two variables. If the correlation value between two variables is towards one then those two variables are said to be highly positively correlated. If the value is zero, then the variables are not correlated. If the value is -1 then the variables are said to be negatively correlated.

5. MODEL BUILDING

5.1. LOGISTIC REGRESSION

Logistic Regression is a very popular statistical algorithm, widely used when the dependent variable is dichotomous. In the telecom data set, the variables are dichotomous since they represent the status of a given customer. More specifically, they highlight the probability of a subscriber to churn in the future. Logistic regression can estimate the probability of the occurrence of an event. Below is the mathematical formula for this model:

$$P(y/x)=1/(1+e^{y})$$

Given the variable x, the conditional probability of the event y is p(y/x), where y is a linear function of the independent input variables:

$$Y=b_o+b_1X_1+b_2X_{2+} \ldots\ldots\ldots$$

The values of (b_0, b_1, b_2) are obtained using the maximum likelihood method. Logistic Regression can be easily applied after data preprocessing and cleaning leading to a quite good performance, also this method is highly effective for producing a binary classification (in our case for predicting if the customer will churn or not).

5.2. RANDOM FOREST

Random forest is an ensemble techniquewhich internally takes mode of all classes and output it as the predicted class in the case of classification and it gives the mean of the classes as output in the case of regression. The optimal decision tree is selected amongst the many and hence it can be a better approach than decision tree alone. Random forest resolves over fitting and it is one of thebest approaches for the churn prediction.

5.3. SUPPORT VECTOR CLASSIFIER

The objective of the Support Vector Machine or SVM is to separate two groups of data points with boundary and the maximize the distance between the two data sets and that boundary. This boundary is a hyperplane since in order to split any N dimensional object into two complete sets you must use an N-1dimensional plane. Unfortunately, SVM is very difficult to visualize beyond 3 features represented on a 3D graph and even harder to attempt to explain. Training time for an SVM is high if any new feature is added. That is the more features you are working with the slower an SVM performs on train time. However, SVM models tend to have good accuracy when they are trained but usually not stand out enough to justify the train time on large feature data sets.

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 2006 - 2012

5.4. XGBOOST

XGBoost is a scalable system for tree boosting. Among many winning solutions in the machine learning competitions in Kaggle, majority of the algorithm used is always XGBoost since 2016.Gradient boosting is the original model of XGBoost, combining weak base learning models into a stronger learner in an iterative fashion. At each iteration of gradient boosting, the residual will be used to correct the previous predictor that the specified loss function can be optimized. As an improvement, regularization is added to the loss function to establish the objective function in XGBoost measuring themodel performance, which is given by

$$J(\Theta) = L(\Theta) + \Omega(\Theta)$$

6. EVALUATION METRICS

In this paper, we consider accuracy, precision, recall, and F-measure as the methods of evaluation to examine the performance of different prediction models. Table 1 shows the confusion matrix in order to calculate these evaluation measures.

6.1. ACCURACY

It is the proportion of the total number of predictions that were correct and is calculated from the equation

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

		Actual Values: Positive (1)	Actual Values: Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Table1: Confusion Matrix

6.2. PRECISION

Table 1 Confusion Matrix,It is the proportion of the predicted positive cases that were correct and is calculated from the equation,

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

6.3. RECALL (or sensitivity)

Recall is the proportion of positive cases that were correctly identified and is calculated from the equation,

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

6.4. F -MEASURE

We cannot rely on only precision or recall to know the performance of the model built. So, F-measure is used to evaluate the performance of the model as it is the harmonic mean of precision and recall.A value closer to one implies that a best combination of precision and recall is achieved by the model.

Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 2006 - 2012

F-Measure = (2 * Precision * Recall)

(Precision + Recall)

Here TP, TN, FP and FN are the True Positive, True Negative, False Positive and False Negative respectively.

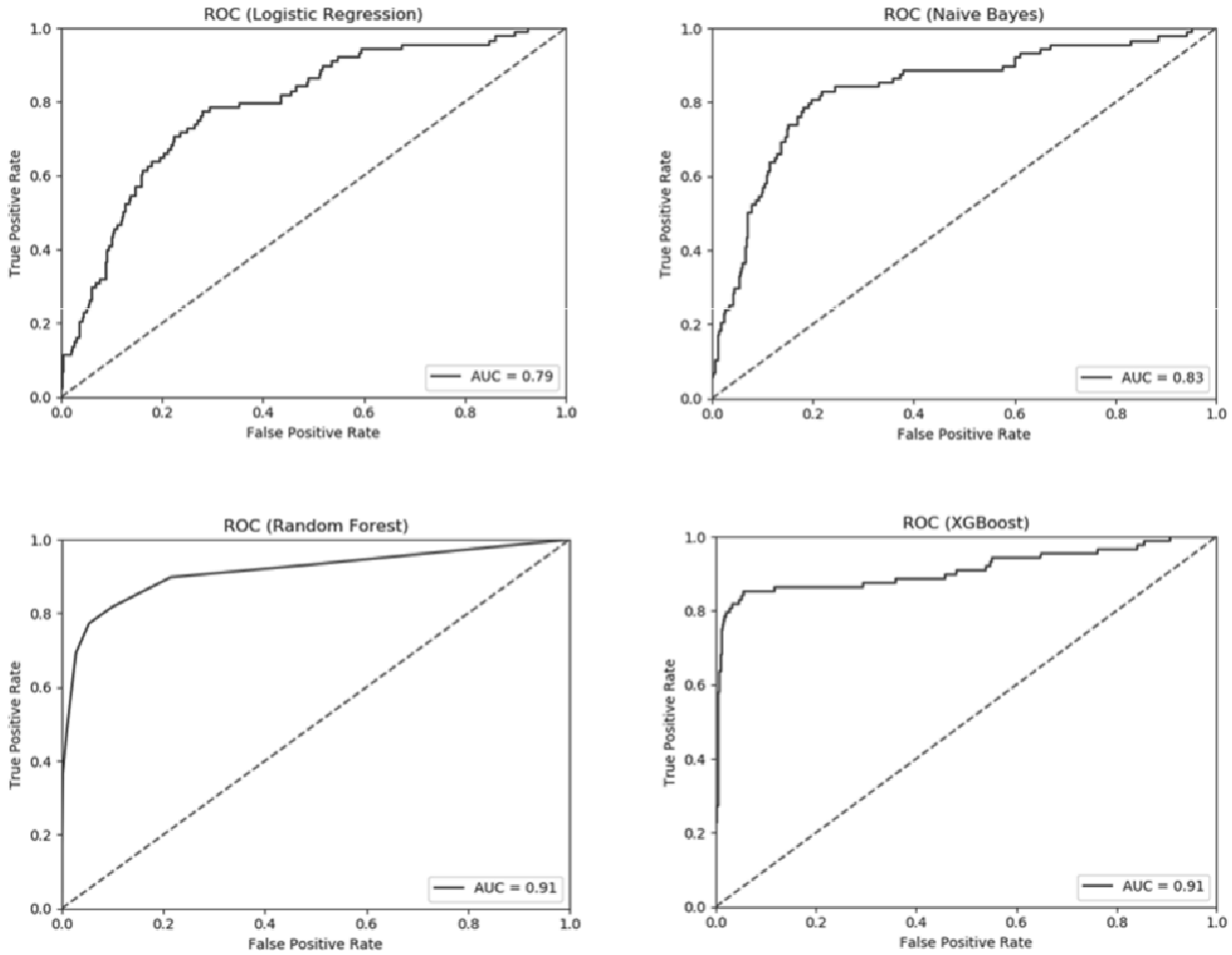
7. RESULT AND DISCUSSION

In this work, the machine learning models are trained on python with several scientific computing libraries, such as NumPy and pandas, which provides efficient data structures and preprocessing methods. Besides, Scikitlearn and XGBoost are imported to support all the learning models.

We have tested all the proposed machine learning algorithms on the original dataset by dividing it into training and test sets. The accuracy of all four classifiers is above 85%, which is a very good result. But if we look at some other metrics like precision and recall, there are some classifiers that have performed better. But to get the best classifier, we must look at every performance metrics. Logistic Regression obtained the lowest accuracy and lowest overall performance, but 85.7% accuracy is still good. XGBoost have obtained the highest accuracy of 95.5% and superior precision, recall and Fmeasure in comparison to all the others.

7.1. Receiver Operating Characteristics (ROC)

Figure 4.4



Published by: The Mattingley Publishing Co., Inc.

March - April 2020
ISSN: 0193-4120 Page No. 2006 - 2012

Figure 4.4 illustrates the ROC curves corresponding to four predictive models. Once again, the Logistic Regression with the AUC (Area Under the Curve) of 0.79, have shown lower performance in comparison to others. Naïve Bayes has a modest AUC of 0.83. Following the curves, one can observe that the ensemble methods provide the best thresholds for separating samples into appropriate classes with the best AUC of 0.91 which is equal for both Random Forest and XGBoost. But still if we have to select just one model, it would be XGBoost because of its high execution speed, performance and its ability to maintain the speed for large data.

8. CONCLUSIONS

In this paper, we explained some methodologies for building the classifier models that will predict customer churn for our business case i.e. in telecommunication sector. The prediction results can be used to build strategies and policies for customer retention targeting the high-risk customers. Furthermore, we observed that for predicting both churners and non-churners, the models have an overall accuracy of 85.7% for LR, 85.1% for NB, 92.3% for Random Forest and 95.5% for XGBoost. In our case, XGBoost worked extra-ordinary in terms of accuracy, AUC and speed. There is no 100 % accuracy because the accuracy is limited by the problem itself, in the sense that there is no 100% correlation between information of customer and their decision to churn. So, 95.5% is very close to the best accuracy for this business problem.

REFERENCES

[1] Chih-Ping Wei, I-Tang Chiu,"Turning telecommunications call details to churn prediction: A datamining approach".Expert Systems with applications 23(2002) 103-112.

[2] MiloˇsMiloˇsevi´c, Nenad ˇZivi´c, Igor Andjelkovi´, "Early churn prediction with personalized targeting in mobile social games," Expert Systems With Applications (2017), doi: 10.1016/j.eswa.2017.04.056

[3] Miguel A.P.M. Lejeune, (2001),"Measuring the impact of data mining on churn management", Internet Research, Vol. 11 Iss: 5 pp. 375 - 387

[4] JOHN POLCARI, "An Informative Interpretation of Decision Theory: Scalar Performance Measures for Binary Decisions," Digital Object Identifier 10.1109/ACCESS.2014.2377593