International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

A Comparative Model for Predicting Customer Churn using Supervised Machine Learning

Muchatibaya Adrin[1], David Fadaralika[2]

School of Information Science and Technology
H190010Z[at]hit. ac. zw[1]
dfadaralika[at]hit. ac. zw[2]
Harare Institute of Technology (H. I. T), Ganges Road, Belvedere, Harare, Zimbabwe

Abstract: ***Churn in customers is an important area of concern for a majority of telecommunications companies. The telecoms industry is of special interest since it suffers annual churn rates of up to 30%. Models have been developed to deal with this problem especially when it is found in such an industry that is the telecoms industry which relies on customers that are not contract based. Predictive models therefore become key to better understand customer churn churn. Handling this issue, in this study the author will implement the SEMMA approach to determine the model with the highest possible accuracy, then choose the best model based on percentage accuracy. This project develops a churn prediction model that can help businesses anticipate which customers are most likely to churn. To discover the key causes of customer turnover, it will employ machine learning techniques such as Random Forest Classifier, Decision Trees, Ada Boost Classifier, SGD Classifier, Logistic Regression, K Neighbors Classifier, Cat Boost Classifier and Gradient Boosting Classifier algorithms. The dataset is comprised of customer demographics, service received and the sum total of their charges from the respective company. It is a Kaggle data set with over 21 attribute obtained from more than 7 000 clients.***

Keywords: Churn management; Wireless telecommunication; Data mining; Decision tree; neural network, big data, Cloud computing

1. Introduction

Historic data is the precursor for the prediction of churn. Churn data from churned clients (response) is considered and their attributes (predictors) that lead to churn. The existing customer's response is predicted by using a statistical model that relates the response to the predictor.

2. Current System

Telecom companies have utilized some methods to predict churn. These methods have utilized data mining and machine learning. A model was developed in South Asia, with the use of AdaBoost & RF techniques, Neural Networks, fuzzy classifiers and SVM Classifier to predict customer churn from a real time data set. Data mining and neural network were also employed in a CRM framework to predict customer behavior in the banking sector. This project utilizes the SEMMA model.

2.1 Classification

Classification determines the category to which a data point (customer) belongs to.

In this case historical or pre-existing data with labels (churner/no-churner) is referred to for use.

The following questions can be answered with classification

Is the customer going to churn?
Will the subscription be renewed?
Is the customer going to downgrade?
Is the behaviour of the user unusual?

2.2 Data collection

The data sources for further predictive modeling can be decided upon once the kinds of insights to look for are identified. This project makes use of a data set with the demographics of the clients, total charges, and service type subscribed to. The data is generated from 7000+ clients distributed over 21 characteristics. Relevant customer data can be analyzed for the development of retention programs.

The following is included in the data set:

Churn record within the last month (Churn services)
Internet, multiple lines, phone
Comprehensive customer information (Contract, mode of payment, services)
Customer demographics

3. Modeling and Testing

A churn prediction model will be developed in this stage. Various models are trained, tuned, evaluated, and tested to determine the one that identifies possible churn with the highest possible accuracy level on evaluation data. Below is a list of the learning models used:

4. Selected Models

1 Logistic Regression
2 Random Forests
3 Decision Trees
4 AdaBoostClassifier
5 GradientBoostingClassifier
6 CatboostClassifier
7 SGDClassifier
8 KNeighborsClassifier

5. Software Requirements

Language: python

Operating System: windows/Linux/ macOS

Tools

Anaconda Navigator
Jupyter Notebook
Numpy
Pandas
Matplotib
Ploty

6. Hardware Requirements

Windows 10 or later operating system
250 GB HDD
2.8GHz Dual Core
4 Gig RAM
Design tools
Scikit learn Machine learning package for python
Python
HTML5
CSS
JavaScript
Visual code IDE and Jupiter note book
MYSQLite or postgres database

7. Requirements Analysis

Historic data is the precursor for prediction of churn. Churn data from churned clients (response) is considered and their attributes (predictors) that lead to churn. The existing customer's response is predicted by using a statistical model that relates the response to the predictor. To predict customer churn is the scope of this work and the process may be broken into the following stages:

The understanding of the problem and the aim

- Collection of data
- Preprocessing and preparation of data
- Testing and Modeling
- Model deployment and monitoring
- Problem understanding and final aim

8. Interface Requirements

This is a boundary on which separate computer components intersect and exchange information. The exchange in this study is between software and hardware, individual users, web applications, end users and combinations of these. The flawless interfacing of information from one system with an additional system prevents information inexactness and surges the susceptibility to hacking into backend systems. This necessitates the creation of secure coding practices through strictly identifying the source of mandatory information, recognizing the information objects and information structures compulsory for the exchange. The clear identification of applicable procedures for the exchange of information and referencing the methodological necessities is key to an interactive interface with high utility.

Software Tools

8.1 Python

Python is a scripting language that is object oriented, interpreted and interactive at a high level. It is highly readable and is ideal for the writing of clear logical code for small to large scale projects. Python is well suited to the rapid prototyping of complex applications.

It is extensible to C or C++ and interfaces with a great number OS system calls and libraries. Python is garbage collected and typed in a dynamic manner. Several paradigms of programming are supported namely functional, procedural and object oriented programming.

8.2 Pandas

Pandas is a package from Python that provides efficient, expressive and flexible data structures which makes it easy to work with time series and structured data. Pandas developed for practically analyzing real world scenarios at a high level in Python.
Columns that are typed heterogeneously, arbitrary matrix data and other forms of data sets for statistics is ideal for Pandas.

The Pandas data structure is comprised of 1-dimensional series and 2-dimensional data sets and these two can handle the vast majority of typical use cases in a number of fields and disciplines and the data needs not be specifically labelled. It is built on NumPy for the proper integration with scientific computing environments and 3rd party environments.

8.3 Anaconda Navigator

This is a GUI for the desktop to function for the launch of applications. It also enables the environments of conda, channels and packages to be easily managed. This is done in the absence of command line commands. Anaconda cloud packages or those in a local repository can be searched for with relative ease and it is available across all major operating systems.

8.4 Executing the code with Navigator

Spyder is the simplest way to it. You click Spyder from the Navigator home and then type your code for execution.

A Jupyter Notebooks can also be used in the same way. They create a web browser viewable notebook file composed of interactive interfaces, output, code and images.

8.5 The Jupyter Notebook

This is a web application that is open sourced which enables sharing of documents created with live code, narrative text, visualizations and equations.

9. Context Level DFD

Activity Diagram

Figure 1: Activity diagram for implementation of Churn Prediction algorithm

Execution flow from the gathered data set to the algorithm training, evaluation of features and result analysis is shown in Fig 1.

10. DFD Level 1

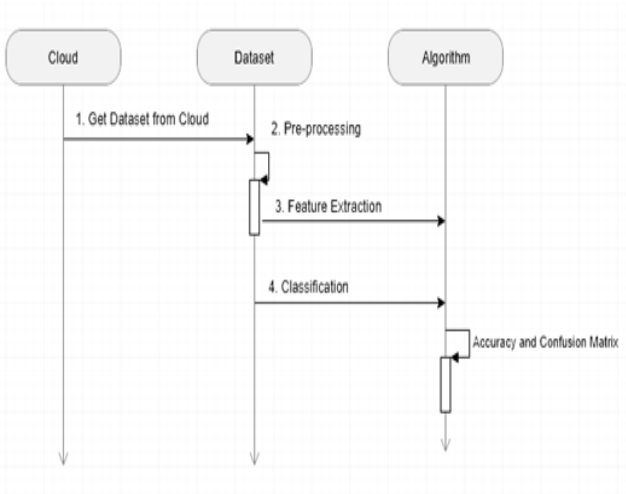

Figure 2: Sequence Diagram for Customer Churn Prediction

11. Performance and Evaluation

The algorithms used were comprehensively evaluated for their respective levels of performance and this can be summarized as follows:

Accuracy score: Generates the model accuracy test and training test.

ROC Curve: The combination of FPR and TPR for different class prediction thresholds to show the diagnostic ability of a model.

AUC for ROC: The measurement of the separability of classes in a model in relation to the ROC curve.

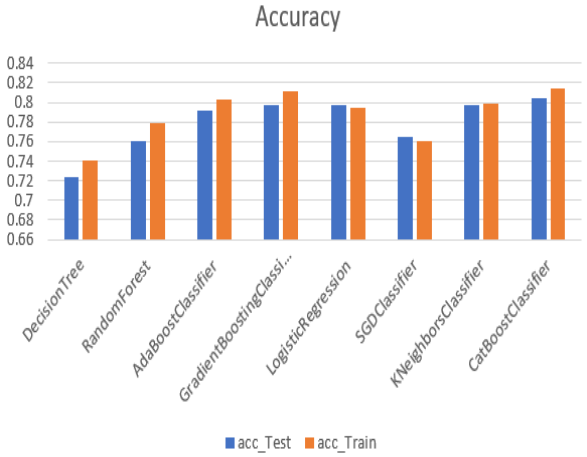

Figure 3: Graphical summarization of the performances of all the algorithms used

Table 1: Comparison of Results

	_ROC_AUC_Test_Data	_ROC_AUC_Train_Full	acc_Test	acc_Train	fit time	model_name
0	0.825115	0.846121	0.723918	0.740149	0.007977	DecisionTree
1	0.843566	0.878392	0.760823	0.778843	0.163078	RandomForest
2	0.839828	0.852390	0.791341	0.803159	0.094746	AdaBoostClassifier
3	0.839062	0.863244	0.797729	0.811502	0.380160	GradientBoostingClassifier
4	0.838241	0.839784	0.797019	0.794995	0.048902	LogisticRegression
5	0.835256	0.837419	0.764372	0.761271	0.017535	SGDClassifier
6	0.835035	0.844151	0.797019	0.798190	0.000000	KNeighborsClassifier
7	0.844791	0.866044	0.804826	0.814341	12.489031	CatBoostClassifier

It can be noted from the table above that the catboost classifier gave the most efficient prediction results evident from the high accuracy, ROC, and AUC values. The model is then deployed because of its satisfactory performance.

11.1 Hyperparameter tuning for CatBoostClassifier

In this study, the hyperparameters of CatBoostClassifier were tuned manually by changing values and running the model. The number of our possibilities is sufficiently small therefore manual hyperparameter tuning became an ideal choice.

11.2 CatBoostClassifier parameters

```
CatBoostClassifier(
    iterations=None,
    learning_rate=None,
    depth=None,
    l2_leaf_reg=None,
    model_size_reg=None,...
    max_depth=None,
    n_estimators=None,
```

11.3 Manual hyperparameter depth tuning

```
#import libraries
from sklearn import datasets
from catboost import CatBoostClassifier
from sklearn.model_selection import cross_val_score

# load data
churn = datasets.WA_Fn-UseC_-Telco-Customer-Churn.csv()

# target
y = churn.target

# features
X = churn.data

#Instantiate CatBoostClassifier, using a maximum depth of 3
cbc = CatBoostClassifier(max_depth=3)

# 5 folds, scored on accuracy
cvs = cross_val_score(cbc, X, y, cv=5, scoring='accuracy')

#Mean value of cross validation score
print (f 'The mean value of cross val score is {cvs.mean()}') #mean = 0.96

print("=======*5)

#Instantiate CatBoostClassifier, using a maximum depth of 5
cbcl = CatBoostClassifier(max_depth=5)

# 5 folds, scored on accuracy
cvsl = cross_val_score(cbcl, X, y, cv=5, scoring='accuracy')

#Mean value of cross validation score
print (f 'The mean value of cross val score is {cvsl.mean()}') #mean = 0.97
```

From this, we can observe that the performance of the model increases with an increase in the maximum depth of results. After the manual tuning the CatBoostClassifier models performance improved from an accuracy level of 86% to 91%.

Below are the dashboard results for the CatBoost model.

12. Assumptions

The system will only analyse churn in customers.

It will predict customer churn classifiers.

13. Conclusion

The study utilized model data from IBM obtained through Kaggle and as much as this data set sufficed to train and test customer churn prediction models an even more accurate and fine-tuned model can be derived from a Zimbabwean data source. Customer churn will continue to be a grievous issue for the telecoms industry therefore a further study is required which will utilize local data to generate customer churn management strategies or at least guidelines to assist telecom companies to better manage customer churn and improve on retention of loyal customers.

References

[1, 2] Berson, A., Smith, S., & Thearling, K. (2000). Building data mining applications for CRM. New York, NY: McGraw-Hill.
[3] “Data Mining: Practical Machine Learning Tools and Techniques, Second Edition-Data Mining Practical Machine Learning Tools and Techniques-WEKA.pdf. ”
[4, 5] H. Ren, Y. Zheng, Y. Wu, Clustering Analysis of Telecommunication Customers. The Journal of China Universities of Post and Telecommunications.16 (2), 114-116 (2009).
[6] K. Dahiya and S. Bhatia, “Customer churn analysis in telecom industry, ” in 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 2015, pp.1–6.
[7] L. Bin, S. Peiji, and L. Juan, “Customer Churn Prediction Based on the Decision Tree in Personal Handyphone System Service, ” in 2007 International Conference on Service Systems and Service Management, 2007, pp.1–5.
[8, 9] M. J. Perez and W. T. Flannery, “A study of the relationships between service failures and customer churn in a telecommunications environment, ” in PICMET ’09-2009 Portland International Conference on Management of Engineering Technology, 2009, pp.3334–3342.
[10] Hemanth Kumar Ravuri, “Study of user’s data volume as function of Quality of Experience for churn prediction” [M. Sc. E. E. thesis 2016: 32, BTH, submitted].
[11] Mounika Reddy Chandiri, “Churn predictive heuristics from Telecom operator and Users' perspective” [M. Sc. E. E. thesis 2016: 05, BTH, submitted].