ISSN - 2348-2397
APPROVED UGC CARE

SHODH SARITA

SHODH SARITA
Vol. 7, Issue 26, April-June, 2020
Page Nos. 177-181

AN INTERNATIONAL BILINGUAL PEER REVIEWED REFEREED RESEARCH JOURNAL

PYTHON : AN EFFECTIVE TOOL FOR DECISION MAKING – A STUDY OF CUSTOMER CHURN IN TELECOM SECTOR

❒ Dr. Jayalekshmi K.R.*

ABSTRACT

Python is one of the rapidly -growing programming languages in the world. Python is used by data scientists to simplify complex data sets and to generate information from a large collection of data. Python enables data scientists to build efficient solutions for complex business problems. It gives huge number of options to data scientists. Python is supported with huge community base where people can exchange their thoughts, ask questions give answer etc. In this paper python is used as a data analysis tool for identifying the key customers who are making their mind to change their mobile network service provider. Two major problems faced by all businesses are customer acquisition and retention. The wireless telephony market is a fast growing service segment. Today majority of phone calls all over the world can be made by mobile phone. Now the mode of struggle got changed from attracting new customers to retaining of old customers. So the companies should have the awareness of customers who are making their mind to stop doing business with their current operator and interested to join with some other of their choice. This is the obtained by churn prediction models. Customer churn is a very important problem which most companies are facing. The industries where the switching cost is too less , the intensity of churn is too high. Telecom sector is very badly affected by this issue. If the service providers use such efficient tools they can reduce the attrition rate of customers and can focus on targeted promotion and retention strategies.

Keywords : Python, data science, Churn prediction

INTRODUCTION

In today's competitive world, customer churn is one of the most critical apprehension for all the mobile network operators . In their endeavor to expand their customer base, or at least maintain the number of customers at a constant level, the providers must stay on their toes in a fierce competition for new customers. The companies can take corrective action to minimize this phenomenon if they are able to recognize the major reasons for the dissatisfaction of clients and to forecast in advance, the clients whom they will lose in near future . It is more expensive for the companies to look for new customer than satisfying an existing customer. Also one bad customer can ruin the likelihood of getting some good customers.

In this industry customers can decide about what quality of service they should receive from their current service provider. If the service provider could not satisfy their clients they can choose any alternate option where they get good quality services. In this highly competitive market, customers will stay with companies who care for them and offers better products and services at lesser prices. Therefore retaining existing subscribers is more important than looking for new ones. Many operators give prime concern for retaining high profitable customers and consider it as the number one business pain. Network service providers make use of effective promotion and customer communication policies to keep their customers happy and force them to stay. For the present scenario retention strategies and churn reduction

*Associate Professor - NCRD's Sterling Institute of Management Studies, Nerul, Navi Mumbai

strategies should be kept as one of major business goal by all the companies. To manage this phenomenon companies should be able to recognize the clients planning for switching and approximate time of their switching .If this is known in advance companies can plan better retention strategies to stop as many customers as possible from switching .

This study is intended to identify the customers who are making their mind to switch the current service provider by predictive modeling with the help of Python. Such Forecasting models helps the companies to recognize in advance which customers are having high chance to change the service provider , why they are changing and when they will churn, improve the quality of services by identifying the areas where improvements are required, provide incentives to the targeted customers, and thereby avoid the economical wastage for mass marketing approaches.

OBJECTIVE

The important aim and objective of this research is to utilize Python to recognize the key customers who are making their mind to switch the current service provider and to develop a proficient and effective model which can spot in advance the probable customers who have made their decision to change the present operator and join a new operator of their choice, in Pre-paid mobile telephony market. This is very much useful for the present carrier to recognize the customers who are thinking about changing their present service provider.

The churn prediction solution uses information about the historical behavior of your customers, revenue, operations, social behavior and other current measures, and applies predictive models to determine the likelihood for churn and build target campaigns towards customer retention. Companies can know what are the important factors contributing the customer's switching decision.

RELATED WORK

Ullah et al.: Churn Prediction Model Using RF , IEEE Access, VOLUME 7, 2019, proposes a churn prediction model that uses classification, as well as, clustering techniques to identify the churn customers and provides the factors behind the churning of customers in the telecom sector. The study first classified customers data using classification algorithms, in which the Random Forest (RF) algorithm performed well with 88.63% correctly classified samples. The model is evaluated using metrics, such as accuracy, precision, recall, f-measure, and receiving operating characteristics (ROC) area. A neural network based methodology for the prediction of churn customers in the telecom sector is provided in [11]. Predictive models for churn customers regarding prepaid mobile phone companies are described in [13]. In another study, authors use Support Vector Machine (SVM), Neural net, Naïve Bayes, K-nearest neighbors and Minimum-Redundancy Maximum Relevancy (MRMR) features selection technique [9]

Churn prediction has been performed in the literature using various techniques including machine learning, data mining, and hybrid techniques. These techniques support companies to identify, predict and retain churning customers, help in decision making and CRM. The decision trees are the most commonly recognized methods used for prediction of problems associated with the customer churn.

RESEARCH METHODOLOGY

This study is intended to find the key customers who are about to churn in telecom sector and to device a predictive model to deal with it. The study is performed on secondary data collected for a specific service provider. By using the tool python, different classifier models are created and the performance evaluation of the models are done, which will in turn will help the service provider to make quality decisions. This process will also help the service provider to predict the class of instances whose class labels are not known.

DATA PREPROCESSING

It is very important for making the data useful because noisy data can lead to poor results. In telecom dataset, there are a lot of missing values, incorrect values like "Null" and imbalance attributes in the dataset. In our dataset, the number of features is 29. We analyzed the dataset for filtering and reduced the number of features so that it contains only useful features. A number of features are filtered using the delimiter function in Python.

The python command isnull().sum() used to view

the missing values of the attribute in the data set. By using the python functions data.head(5) and "data.shape" can be used to get general view of the dataset. The categorical values are converted to numerical values for making the analysis easy by the alogorithm. The replace function is used here for it. Also some irrelevant attributes which are not used in the model are also eliminated remove the columns not used in the predictive model.

```
data['international plan'].replace('yes',1, inplace=True)
data['international plan'].replace('no',0, inplace=True)
data['voice mail plan'].replace('yes',1, inplace=True)
data['voice mail plan'].replace('no',0, inplace=True)
```

After preprocessing the data set consists of 3333 records with 18 attributes of which 10 are integer type and 8 are float.

```
account length              3333 non-null int64
international plan          3333 non-null int64
voice mail plan             3333 non-null int64
number vmail messages       3333 non-null int64
total day minutes           3333 non-null float64
total day calls             3333 non-null int64
total day charge            3333 non-null float64
total eve minutes           3333 non-null float64
total eve calls             3333 non-null int64
total eve charge            3333 non-null float64
total night minutes         3333 non-null float64
total night calls           3333 non-null int64
total night charge          3333 non-null float64
total intl minutes          3333 non-null float64
total intl calls            3333 non-null int64
total intl charge           3333 non-null float64
customer service calls      3333 non-null int64
churn                       3333 non-null int64
dtypes: float64(8), int64(10)
```

FEATURE SELECTION

After the data preprocessing the correlation analysis between the churn and each customer feature is done to determine which features to be included in the prediction model.

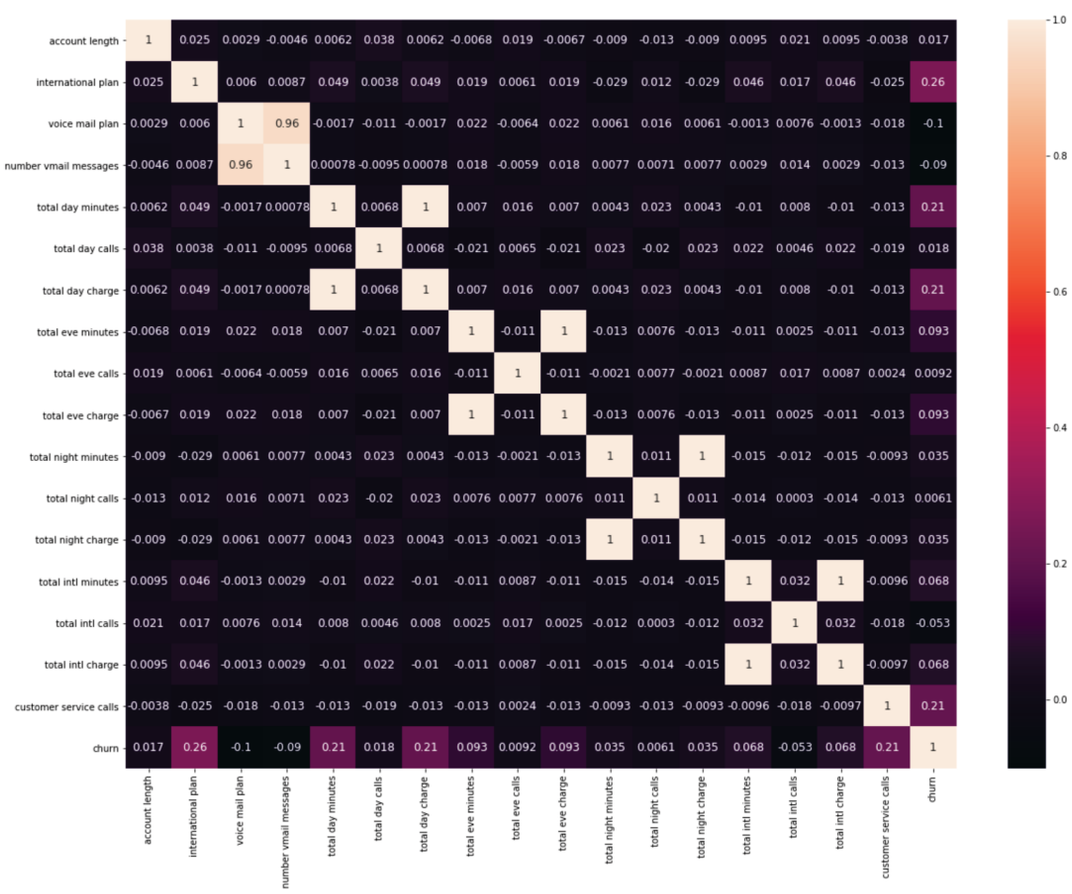

Table-1

From the correlation matrix the highly correlated customer features to the target feature churn are identified and are used in the predictive modeling.

PREDICTIVE MODELING

Different models are considered for predicting the customer churn. The churn prediction models are created by using Logistic regression, Random forest, J48 and Naïve bayes.

The data set of 3333 records are split into training and test set with tests set consisting of 25% of the data and training set with 75% data.

- **Logistic regression**

Logistic regression is a basic classification algorithm. It is generally used for binary classification problem. The output of the logistic regression is a probability.

The dependent variable is categorical here. By using python Spyder the code for logistic regression model is created .

The model is imported from Sklearn. After training the model on the training data set, we can use this model on the test data set. The results are saved in the variable "prediction_test1" and later the accuracy score is measured and printed.

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
result = model.fit(X_train, y_train)
from sklearn import metrics
prediction_test1 = model.predict(X_test)# Print the prediction accuracy
print (metrics.accuracy_score(y_test, prediction_test1))
```

- **Random Forest**

```
from sklearn.ensemble import RandomForestClassifier
randomForest1 = RandomForestClassifier()
randomForest1.fit(train_x, train_y)
print('Accuracy of random forest classifier on test set: {:.2f}'.format(randomForest1.score(test_x, test_y)))
```

Similarly the predictive modeling using J48, and Naïve Bayes are also done .

PERFORMANCE EVALUATION

The prediction model is evaluated by accuracy, precision, recall and F-measure. Accuracy of the model indicates the instances that are correctly classified . Accuracy is calculated as

Accuracy= (TP+TN)/ (TP+TN+FP+FN)

TP indicates the True positive ,TN True negative, FP False positive, FN False negative. The TP rate of the model indicates the portion of data that is correctly classified as positive.

This measure is also called as Sensitivity.For any model it is highly desirable to have a high value for TP Rate and low value for FP Rate. TP Rate is calculated as

TP Rate= TP /Atual Positives

The FP Rate indicates the portion of data which are incorrectly classified as positive. It is calculated as

FP Rate = F P / Actual Negatives

The next measure used is Precision. The precision tells about what percentage of data tuples the classifier labeled as positive are actually positive . The accuracy is also known as positive predictive value(PPV). It is calculated as

Precision= TP / (TP+ FP)

Another measure used is Recall .The recall is the ratio of correctly predicted positive values to the actual positive values .it is calculated as

Recall= TP / (TP + F N)

RESULTS & FINDINGS

The results shows that J48 has the highest value for recall.

It indicates that the algorithm could find the maximum number of true positives in the data set and can correctly identify the true churners. The next good performer is Random forest with the second highest value for recall.

Similarly the precision values for J48 and Random forest is high as compared to the other algorithms .

This indicates that J48 and Random forest outperform the other algorithms in predicting the real positive values.

Also the TP rate of J48 and Random forest is also high . More over the F-measure of J48 and Random Forest is high compared to the other classifiers. The F-measure for J48 is 91% and that for Random forest is 87.6%.

Method	TP rate	FPrate	Precision	Recall	F-measure
Logistic regression	0.853	0.76	0.81	0.84	0.81
Random Forest	0.896	0.55	0.89	0.895	0.876
J48	0.90	0.41	0.91	0.912	0.905
Naïve Bayes	0.88	0.53	0.86	0.88	0.87

Table-2

CONCLUSION

The major crisis that the service provider was facing was to identify the future churners and to focus them with incentives so that they are convinced to stay back. Due to the absence of a precise model to monitor the customer behavior, the company was not capable distinguish the churners from non-churners.

To address this problem the predictive models that are created and they proved competent enough to make out churners and non-churners. This will help the service provider to carry out well-organized retention campaigns. This has become a resourceful manifestation to reduce the cost of marketing and rate of churn. As the values of precision, recall, TP rate and F-measure are higher for J48 and Random Forest, these classifiers could identify the maximum customers who are planning to change the service providers form the data set. So for the data set under study, the J48 and Random Forest performed well for identifying the true churners.

FUTURE ENHANCEMENT

In this study only four classifiers are constructed and their performance evaluation is done. More classifiers can be constructed by using Python and the performance evaluation can be done . Even the pictorial representations of the results are not included here. Python's libraries can be used to depict the results .

References :

1. C. Geppert, "Customer churn management: Retaining high-margin customers with customer relationship management techniques," KPMG & Associates Yarhands Dissou Arthur/Kwaku Ahenkrah/David Asamoah, 2002.
2. [2] Y. Huang, B. Huang, and M.-T. Kechadi, "A rule-based method for customer churn prediction in telecommunication services," in Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining. Berlin, Germany: Springer, 2011, pp. 411–422.
3. A. Idris and A. Khan, "Customer churn prediction for telecommunication: Employing various feature's selection techniques and tree based ensemble classifiers," in Proc. 15th Int. Multitopic Conf., Dec. 2012, pp. 23–27.
4. D. Manzano-Machob, "The architecture of a churn prediction system based on stream mining," in Proc. Artif. Intell. Res. Develop., 16th Int. Conf. Catalan Assoc. Artif. Intell., vol. 256, Oct. 2013, p. 157.
5. P. T. Kotler, Marketing Management: Analysis, Planning, Implementation and Control. London, U.K.: Prentice-Hall, 1994.
6. A. Sharma and P. K. Kumar. (Sep. 2013). "A neural network based approach for predicting customer churn in cellular network services." [Online]. Available: https://arxiv.org/abs/1309.3945