A COMPARATIVE ANALYSIS OF PREDICTIVE MODELS USING MACHINE LEARNING ALGORITHMS FOR CUSTOMER ATTRITION IN THE MOBILE TELECOM SECTOR

Dr.Jayalekshmi K.R, Assistant Professor
NCRD's Sterling Institute of Management Studies
Navi Mumbai
dr.lekshmijaya2005@gmail.com

Dr. Jyoti Kharade, Vice Principal & Associate Professor
Bharati Vidyapeeth's Institute of Management and Information Technology
Navi Mumbai,
jyoti.kharade@bharatividyapeeth.edu

ABSTRACT
Machine learning (ML)is important in data analysis and decision-making for any business.ML algorithms are used to analyze complex data sets and to extract useful insights from a large collection of data. It enhances the efficiency of decision-makers to arrive at better solutions for complex business problems. Two major glitches faced by all businesses are customer acquisition and preservation. The very first thing the companies can do to reduce the attrition rate is to understand their customers. Even if some customers have churned, the companies should analyze the reasons for their attrition, so that they can make use of this information to reduce the future attrition rate. In this paper, different algorithms are used for identifying the key customers who are making up their minds to switch their mobile network service provider. Predictive modeling with KNN, SVC, logistic regression, decision tree, and random forest with its performance evaluation for prediction of customer attrition. 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: Customer attrition, ML, data analysis, logistic regression, SVC.

Introduction

Issues like churn of a customer to a competitor is faced by almost every organization in any field. It leads to a major source of financial loss as it is normally more expensive to generate new customers over the retention of the existing ones. It is therefore important to manage churn, where strong competition and saturated markets prevail in industries such as mobile telecom. Without an efficient predictive model, proactive churn and retention will not give the desired outcome. Such a tool will enable us to identify which customers to contact proactively, in order to avoid churn. Hence churn forecasting models plays a vital role in the survival of telecommunication service providers. 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. One bad customer can also ruin the likelihood of getting some good customers.

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 of churn and build target campaigns toward customer retention.

The research paper intends to mark customers who are likely to switch their current service provider and join another company by predictive modeling using different ML algorithms and comparative analysis of their performance a model is built. Such Forecasting models help the companies to recognize in advance the probable customers who will change their service provider, why they are changing, and when they will churn. They can also be used to improve the excellence of services by identifying the areas where enhancements are required, providing incentives to the targeted customers, and thereby planning cost-effective marketing strategies.

Objectives of the study

The following are the objectives of the paper:
1) To create predictive models using different ML algorithms.
2) To measure the accuracy of the created forecasting models in predicting future churners.

Literature Review

Wolniewicz.& Dodier (2004), Customer churning is the key concern in telecommunication. Most organizations are willing to have a long-term relationship with their customers hence their needs and behavior are required to be understood properly from time to time. To predict churn effectively different feature engineering techniques to get the hidden relationships between different entities of the database are studied.

Chawla (2005), used churn analysis as a method to predict customers who are likely to quit the use of a product or service. Basiri, Taghiyareh & Moshiri (2010) hybrid approach is created using an ordered weighted average to add the output of each learned classifier and improve result accuracy. Bagging and boosting are used to train the classifiers. The approach was pretty good compared with many known classifiers.

Umayaparvathi & Iyakutti,K. (2012), churn prediction through ANNs shows better accuracy over decision trees. Dalvi, Khandge, Deomore, Bankar& Kanade (2016), ML techniques like logistics regression and decision trees were used to build the model to predict customer churn in the telecom sector.

Utku, Candan, & Ince (2012) ML techniques used and performance analysis for better prediction and accuracy did using the WEKA tool and ensemble methods showed random forest to perform best with 0.6533 with 50 trees depth.Ullah, Raza, Malik, Imran, Islam & Kim (2019) analyzed various customer churn patterns via performing comparative analysis on techniques such as random forest, SVM, Extreme Gradient Boost (XGBoost), ridge classifier, and neural networks. Mishra, Reddy, & Srinivasulu (2017) built models and compared them with Naive Bayes Classifier, Decision Tree, and Support Vector Machine (SVM) which showed that ensemble-based classifiers had less error rate, low specificity, high sensitivity, and greater accuracy.

Khalid, Mokhairi,& Abdul,(2017), classification models based on Bayes Net, Simple Logistic, and Decision Table with two feature reduction algorithms like correlation-based feature selection (CFS) and Information Gain (IG) built and the result shows performance get improved of the classifiers with use of features reduction of customer churn data set. Au, Chan &Yao (2003), Multi-layer perceptron supervised model trained with Back Propagation algorithm (BPN). The neural network achieves better performance over decision trees.

Materials and Methods

The performance of different ML models that are created is analyzed to identify customers who are about to change their current service provider. The study is performed on open-source data collected from Kaggle. Five ML algorithms are used for the predictive modeling and the performance evaluation of the models is done. The algorithms used for model building are Logistic regression, SVC, Random Forest Decision Tree, and KNN. The implementation part is done by using Python. Comparative analysis of the machine learning models is done that is created to forecast customer attrition. The below figure shows the framework of the predictive modeling.

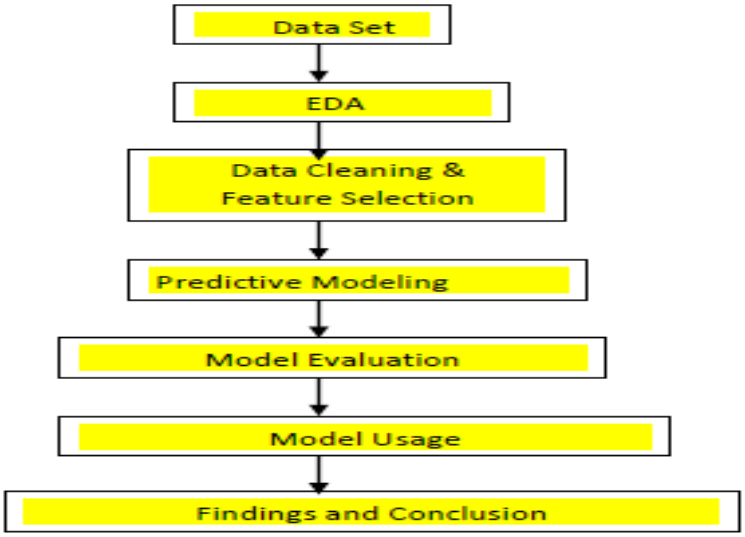

Figure No.1: Framework of the predictive modeling (Source-compiled by researcher)

The data set contains 7043 records with 21 attributes. Exploratory data analysis is performed on the data set.
data. shape

(7043, 21)

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
customerID          7043 non-null object
gender              7043 non-null object
SeniorCitizen       7043 non-null int64
Partner             7043 non-null object
Dependents          7043 non-null object
tenure              7043 non-null int64
PhoneService        7043 non-null object
MultipleLines       7043 non-null object
InternetService     7043 non-null object
OnlineSecurity      7043 non-null object
OnlineBackup        7043 non-null object
DeviceProtection    7043 non-null object
TechSupport         7043 non-null object
StreamingTV         7043 non-null object
StreamingMovies     7043 non-null object
Contract            7043 non-null object
PaperlessBilling    7043 non-null object
PaymentMethod       7043 non-null object
MonthlyCharges      7043 non-null float64
TotalCharges        7043 non-null object
Churn               7043 non-null object
dtypes: float64(1), int64(2), object(18)
```

Table1:Dataset information

After this, the data preprocessing step is done as it is the most time-consuming and most important step in model building. The data set may have incomplete, inconsistent, and noisy data, so this is the most important and mandatory step. Missing, Null values, and imbalance attributes in the dataset are treated, and unimportant attributes are removed.

```
In [10]: data.isnull().sum()

Out[10]: customerID          0
         gender              0
         SeniorCitizen       0
         Partner             0
         Dependents          0
         tenure              0
         PhoneService        0
         MultipleLines       0
         InternetService     0
         OnlineSecurity      0
         OnlineBackup        0
         DeviceProtection    0
         TechSupport         0
         StreamingTV         0
         StreamingMovies     0
         Contract            0
         PaperlessBilling    0
         PaymentMethod       0
         MonthlyCharges      0
         TotalCharges        0
         Churn               0
         dtype: int64
```

Table2: Dataset missing values count

From the above function, it is very clear that the data set contains no missing values for any of the attributes. But the attribute ‘Total Charges’ which is in the object data type needs to be converted to numeric type and checked for missing values.

```
data['TotalCharges ']= pd.to_numeric(data.TotalCharges,errors='coerce')
data.isnull().sum()
```


Above command, after execution shows that attribute "Total Charges" generated 11 missing values which are replaced by the mean value of the attribute. Some binary categorical attributes were converted into numerical ones. Attribute like'Senior-citizen', Yes", values mapped to 1 and "No" to 0. The data set is ready to take for model building.

Analysis and Findings

After the data preprocessing the features for the model building were identified. The final features count was 20 features and 7032 records. Out of the non-churners, around 49% were females and 51% were males.

```
data["Churn"][data["Churn"]=="No"].groupby(by=data["gender"]).count()
 gender    count
Female    2544
Male      2619
Name: Churn, dtype: int64
```

Similarly, out of the churn class in the data set, 939 were female and 930 were male. A different analysis of the variables is done with the target churn variable. The analysis of customer contracts to target variable showed that maximum customers who had contracts month-to-month have made their minds to move out as compared to customers who had 1 and 2-year contracts. Analysis of customer 'Payment Method', 'Internet service', 'Dependents', 'Partners', 'SeniorCitizen' PaperlessBilling', TechSupport' were also analyzed w.r.t the target variable and could generate the following observations from the data set.

The churn rate of customers who opted for Electronic Checks as a Payment Method was more as compared to other options. Customers using Fiber optic have a high attrition rate which indicates dissatisfaction with internet service. Customers who were using DSL service showed less attrition rate compared to Fibre optic service. Dependents distribution with the target variable revealed that customers without dependents have a high attrition rate. Partners distribution with target variable showed that customers without partners are having high churn rate.

The analysis of the attribute 'SeniorCitizen' with the target variable revealed that also the fraction of 'Senior Citizen' who moved out was very less. The analysis revealed that customers who opted for the 'Paperless-Billing' option are more likely to churn. Similarly, analysis of 'Tech Support' with target variable showed that customers with no 'TechSupport' are more probable churners.

The prediction models are created by using KNN, SVC, logistic regression, decision tree and random forest. The data set of 7032 records are split into training and test set with tests set consisting of 30% of the data and training set with 70% data.

KNN Model: -

```
knneighbor_model1=KNeighborsClassifier (n_neighbors=1)
knneighbor_model1.fit (x_train,y_train)
predicted_y= knneighbor_model1.predict (x_test)
accrcy_knn= knneighbor_model1.score (x_test,y_test)
print("accuracy:", accrcy_knn)
KNNeighbor accuracy: 0.7753554
```

SVC Model: -

```
Svc_model2=SVC (random_state=1)
Svc_model2.fit (x_train, y_train)
Predict_y=svc_model2.predict (x_test)
Accuracy_svc=svc_model2.score (x_test,y_test)
Print ("SVM accuracy is:", Accuracy_svc)
SVC accuracy is: 0.80758
```

Logistic regression model: -

```
lr_model3=LogisticRegression ()
lr_model3.fit (x_train,y_train)
accuracy_lr= lr_model3.score (x_test, y_test)
Print ("Logistic regression accuracy is:", accuracy_lr)
LR accuracy= 0.809005
```


Decision Tree Model: -

```
dtree_model4=DecisionTreeClassifier ()
dtree_model4.fit (x_train,y_train)
predictdtree_y= dtree_model4.predict (x_test)
accuracy_dtree= dtree_model4.score (x_test,y_test)
print (" D Tree accuracy:", accuracy_dtree)
```

D Tree accuracy is: 0.725118

Random Forest: -

```
model5_randomf=RandomForestClassifier(n_estimators=500,oob_score=True,n_jobs=1,
random_state=50, max_features="auto", max_leaf_nodes=30)
model5_randomf.fit (x_train,y_train)
predi_test= model5_randomf.predict (x_test)
print(metrics, accuracy-score(y_test,predi_test))
```

R Forest Accuracy = 0.813744

Model Name	Model Accuracy	Performance
Random Forest	0.813744	Best performing model
Logistic regression	0.809005	Average
SVC	0.807583	Average
KNN	0.775355	Underperforming
Decision Tree Tree	0.725118	Underperforming

Table-3: Performance of the models

Random forest classifier has the highest accuracy of 81.37%. Logistic Regression is next good performer with the second highest value for accuracy of 80.90%, followed by SVC with an accuracy score of 80.75% and KNN classifier with 77.5% and the lowest performer is Decision Tree with an accuracy score of 72.5%.

Conclusion

Customer attrition is definitely a bad phenomenon for any organization's profitability. One of the best things the organization can do is to know its customers by improving at customer gratification and also spotting the customers who are about to switch. This information can be used to provide the customers with some promotional offers and hence to retain them from churning. Also, the reason why some of the customers have churned can also be used to avoid future customer attrition.

The major problem faced by the telecom service providers is to identify the churners and concentrate on them by providing incentives so that they are convinced to stay back. The need for a precise model to monitor customer behavior is felt where without such a tool the companies are not able to differentiate churners from non-churners.

To address this problem the predictive models are created and they proved competent enough to make out churners and non-churners. This will help the service providers to carry out well-organized retention campaigns. This has become a resourceful indicator to reduce the cost of marketing and the rate of churn.

References

Au, W. H., Chan K.C., Yao X. (2003), "A novel evolutionary data mining algorithm with applications to churn prediction", IEEE Trans. Evol. Computation. 7 (6) 532–545.

Khalid, A., Mokhairi, M. Abdul, R. (2017)"Improving Accuracy and Performance of Customer Churn Prediction Using Feature Reduction Algorithms", Journal of Telecommunication, Electronic and Computer Engineering-ISSN: 2289-8131 Vol.9No.2- 3.

Basiri, J., Taghiyareh, F., and Moshiri B.(2010), “A hybrid approach to predict churn,” in Services Computing Conference (APSCC), IEEE Asia Pacific. IEEE, 2010, pp. 485–491.

Chawla N. V. (2005), “Data mining for imbalanced datasets: An overview,” in Data mining and knowledge discovery handbook. Springer, pp. 853–867.

Dalvi, PK., Khandge S.K., Deomore, A., Bankar, A., and Kanade V. A. (2016), “Analysis of customer churn prediction in telecom industry using decision trees and logistic regression”, In Colossal Data Analysis and Networking (CDAN), Symposium on, pp. 1-4. IEEE.

Abinash, M., Reddy, U., Srinivasulu (2017), “A Comparative Study of Customer Churn Prediction in Telecom Industry Using Ensemble Based Classifiers”, Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017) IEEE Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9.

Ullah I, Raza B, Malik A. K, Imran M., Islam S. ul, and Kim S. W. (2019), “A Churn Prediction Model using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector", IEEE.

Umayaparvathi V., and Iyakutti K (2012), "Applications of data mining techniques in telecom churn prediction." International Journal of Computer Applications 42, no. 20 : 5-9.

Wolniewicz R.H., and Dodier,R,(2004), “Predicting customer behavior in telecommunications”, IEEE Intell. Syst. 19 (2) 50–58.

Utku Y., Candan, C., and Ince,T. (2012), "Customer Churn Prediction for Telecom Services", In Computer Software and Applications Conference (COMPSAC), 2012 IEEE 36th Annual, pp. 358-359. IEEE.

