Telecom Customer Churn Prediction Using Machine Learning

Abstract- Telecommunication industry provides customers an opportunity to choose from various service providers. However, certain factors such as low switching costs and deregulation by the government have contributed to the risk of customers switching to competitors. Customer churn can therefore be defined as the switching of a customer from services of one provider to another. Since it can be a costly risk, it needs to be managed properly. The presented paper aims at predicting customer churn using diiferent algorithms like Naive Bayes,Decision Tree,Logistic Regression,K-Nearest Neighbor and Support Vector Machine.

Keywords- Churn Prediction, Classification, Decision tree(DT), Random Forest(RF), Naive Bayes(NB),Logistic Regression(LR),K-Nearest Neighbor(KNN),Support Vector Machine(SVM),Machine Learning(ML),prediction

I. INTRODUCTION

One of the key aspirations of any telecommunication company is to maintain a loyal customer base but since the customers have been provided with facility of switching from one service provider to another, telecommunication companies' are facing more problems. According to a study, acquiring a new customer is about 5-6 times costlier as compared to retaining an old one [1][2]. Customer churn is one of the major issues that the telecom industry is facing today. Hence customer churn may prove to be a costly risk if not managed carefully. Various costs are associated with customer churn and include loss of revenue, costs of customer retention and reacquisition, advertisement costs, organizational as well as planning and budgeting chaos [2]. Therefore it becomes quite necessary to identify the possible churning customers so that the losses can be prevented. Data mining methods can be defined as the process of finding unknown patterns in huge data sets [8]. These methods find their applications in the area of CRM such as fraud detection, customer churn prediction etc. Classification is one of the data mining tasks that focuses on classifying unknown cases based on a set on known examples [9]. Hence, classification techniques based on data mining can be used for predicting churn in telecommunication industry[8][10].Wai-Ho Au et al. [3], Erfaneh and

Tarokh [4], Wei Yu et al. [5], Chao et al. [6], Adnan and Asifullah [7] have also focused on using data mining techniques for churn prediction in telecommunication in their work. Various techniques have been used by various researchers but use of data mining techniques for predicting customers churn has turned out to be an efficient approach with high accuracy in results.

The rest of the paper is organized as follows: In the Second Section, a rich literature survey has been provided to put a spotlight on the related work that has been done in the field of churn prediction. The third section focuses on the data used for research purpose. The fourth section comprises of the algorithms implemented on the dataset. In the next section, the results obtained after implementation have been mentioned and analyzed. Finally the conclusion has been presented with some future scope of work.

II. LITERATURE REVIEW

- Related Work

Clement et al. [2] have presented new features categorized as contract-related, call pattern description, and call pattern changes description features derived from traffic figures and customer profile data. The given features were evaluated using Naïve Bayes and Bayesian network and obtained results were compared to results obtained using decision tree. Results have shown that probabilistic classifiers have shown higher true positive rate than decision tree but decision tree performs better in overall accuracy. Essam et al. in [11] have introduced a simple model based on data mining to track customers and their behavior against churn. A dataset of 500 instances with 23 attributes has been used to test and train the model using 3 different techniques i.e., Decision trees, SVM and Neural networks for classification and k-means for clustering. Results indicate that SVM has been stated as the best suited method for predicting churn in telecom. Umman Tuğba Şimşek Gürsoy [12] have compared regression techniques with decision tree based techniques. Results have shown that in logistic regression analysis churn prediction accuracy is 66% while in case of decision trees the accuracy measured is 71.76%. Hence decision tree based techniques are better to predict customer churn in telecom. V. Umayaparvathi and K. Lyakutti [13] have used Neural Networks and Decision trees to build the churn prediction model. According to the results, Decision trees have 98.88% of predictive accuracy and an error rate of 1.11167%. Similarly neural networks have shown the predictive accuracy of 98.43% with 1.5616% of error rate. As is indicated by the results, decision trees have outperformed neural networks for churn prediction. According to the authors, selection of right combination of attributes and fixing the proper threshold values may produce more accurate results.

Saad et al. [14] have applied different machine learning algorithms such as linear and logistic regression, ANN (Artificial Neural Networks), K-means clustering, Decision Trees to identify churners and active customers. The best results were obtained using exhaustive CHAID, a variant of standard decision trees. Ning Lu [15] has proposed a model with an “Implementation zone” where customers with highest churn probability can be addressed for retentive actions. The author has also proposed a further improvement in performance by analyzing other classification techniques as well or using a hybrid approach for more accurate results. Vladislav and Marius in [17] have presented quality measures of six churn prediction models including regression analysis, naïve Bayes, decision trees, neural networks etc. They have also pointed out the links between churn prediction and customer lifetime value. According to the authors, new prediction models need to be developed and combination of proposed techniques can also be used. Khalida et al. in [20] have used a specific training sample set was used to conduct an experiment on customer churn factor using decision tree. According to the authors, rule information can be easily understood by decision tree. An attempt has been made to identify various factors responsible for customer churn such as area. Amal et al.[21] have reviewed that generally Decision tree based techniques, neural network trees and Regression techniques are applied in churn prediction. Decision tree based techniques outperform all other in terms of accuracy. On the other hand, neural networks outdo other techniques due to size of data sets. From the presented literature work, it can be concluded that in most of the cases, Decision trees have outperformed other techniques for predicting churn in telecommunication industry

III. DATASET

Churn analysis is done on the basis of historical data. This data may be accessible from the warehouse of respective company. The data set used in this study was acquired from an online source1 . This data set is a longstanding customer data set of about 10,000 customers (active and disconnected) and include demographic as well as service details such as their customer id, credit score,estimated salary , etc. A customer is considered as active if he is still using the network and in case the services are terminated either voluntarily or involuntarily [11], the customer is disconnected.

Number	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
9996	15606229	Obijiaku	771	France	Male	39	5	0.00	2	1	0	96270.64	0
9997	15569892	Johnstone	516	France	Male	35	10	57369.61	1	1	1	101699.77	0
9998	15584532	Liu	709	France	Female	36	7	0.00	1	0	1	42085.58	1
9999	15682355	Sabbatini	772	Germany	Male	42	3	75075.31	2	1	0	92888.52	1
10000	15628319	Walker	792	France	Female	28	4	130142.79	1	1	0	38190.78	0

IV. METHODOLOGY

(A) Data Collection

In this research paper we have used Real world data set for predicting churn status of the customers.The data collected is widely used data and is available at UCI Machine Learning Repository. This Real data belongs to Telecom Company. It consists of record of 10,000 people . The data consists of real numbers, Decimal values and Nominal values

(B) Data Pre-Processing

Data pre-processing is a way to convert the noisy and huge data into relevant and clean data, as the data available is Real world data, so it contains inaccurate data, missing values and other Noisy data, for removing this inconsistent data from the Dataset, the proposed system have to clean the raw data.
Following are the step involved in data pre processing:

- Looking Up For Proper Format: As we have made our model using python, so we need a csv file (comma separated value) for our code. The data downloaded is in the form of RAR file, so we extract the data from the text file available and save it into a csv file so that our python code can read it. This is the first most important step, if the data

is not available in requires format then we cannot design the classification model.

.● Finding Missing Values: When the data collected is real world data, and then it will contain missing values. This brings more change in the prediction accuracy. Sometimes these missing values can be simply deleted or ignored if they are not large in number. It is the simplest way to handle the missing data but it is not considered healthy for the model as the missing value can be an important attribute .The missing values can also be replaced by zero this will not bring any change as whole, but this method cannot be much yielding. So an efficient way to handle missing values is to use mean, average of the observed attribute or value. This way we lead to more genuine data and better prediction results

● Data Reduction :It means to reduce the number of features while maintaining a good analytical result.For thi purpose,feature selection and feature associations or correlation have been studied to remove redundant inormation.

(a) Feature Selection:In this step we select subset of relevant attributes from the total given attributes. This stage helps in reducing the dimensionality and making the model simpler and easy to use, thus leading to short training time and high accuracy.
To obtain highly dependent features for Churn prediction we have used Correlation and dependence method. The term correlation can be defined as mutual relationship between two. In this those attributes are chosen which highly influence the occurrence of Churn

(b).Data Transformation: In this step we transform the given real data into required format. The data downloaded consist of Nominal, Real and Decimal values. In this step we convert the Nominal data into numerical data of the form 0 and 1. The positive value is assigned the value of 1 and the negative value is assigned the value of 0. Now the resultant csv file comprises of all the integer.

(C)Modeling

In the modeling stage, six machine learning algorithms
have been applied to the dataset to assess their ability to detect
Churn . These algorithms are logistic regression (LR), support
vector machines (SVM), random forest (RF), decision tree(DT) ,Naive Bayes(NB) and k-nearest neighbors(KNN).

(A) Naïve Bayes:

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions among predictors. A Naive Bayesian model is easy to build, with no complex iterative parameter assessment which makes it especially useful for very large datasets. Even though it's simple, the Naive Bayesian classifier often does unexpectedly well and is widely used because it often outperforms more refined classification methods.

Algorithm

• Bayes theorem delivers a way of calculating the posterior probability, P(c|x), from P(c),P(x) and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This hypothesis is called class conditional independence [16].

(c/x)=p(x/c) p(c)/p(x)

(1) P (c|x) is the posterior probability of given predictor. P(c) is the prior probability of class. P (x|c) is the likelihood which is the ability of predictor given class. P(x) is the prior probability of predictor.

(B)Decision Tree:

Decision tree builds classification or regression models in the form of a tree like structure. It breakdowns a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally established. The last result is a tree with decision nodes and its leaf nodes. Decision nodes have two or more branches. Leaf node represents a classification or decision. The uppermost decision node in a tree which resembles to the finest predictor called root node. Decision trees can switch both categorical and numerical data values. The core algorithm for building decision trees is called ID3 by J. R. Quinlan which employs a top-down and greedy search over the space of possible branches with no backtracking.

Algorithm

• Start with single node N, with training data D.

• If all the data in D belongs to same class, then N becomes leaf. Otherwise attribute 'A' is selection method based on splitting criterion.

• The instance in 'D' is partitioned accordingly

• Apply algorithm recursively to each subset in 'D' to each subset in 'D' to form decision tree.

The algorithm uses Entropy and Information Gain to construct a decision tree [17].

. Entropy ID3 algorithm uses entropy to calculate the similarity of If the sample is

completely similar the entropy is zero and if the sample is an equally divided it has entropy of one.
E(S)=-Pi log2 pi ------(2)
Information Gain
The information gain is based on the decrease in entropy after a dataset is split on an attribute. Creating a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches).
GAIN(T,X)=ENTROPY(T)-ENTROPY(T,X) ––– (3)

(C) K- Nearest Neighbour:

K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique. A case is classified by a majority vote of its neighbours, in the case being assigned to the class most common between its K nearest neighbours measured by a distance function. If K = 1, then the case is simply allocated to the class of its nearest neighbour [18].
Algorithm
• Training set: (x1,y1),(x2,y2),,,,,(xn,,yn). • Assume X:= (x:(1) ,x: (2) ,...,x: (d)) is dimensional feature vector of real numbers for all i
• Y: is a class label {1...C}, for all i.
• Find the closest point Xj to Xnew using distance measures.
• Classify by Y knn= majority vote among the K points.

(D) Support Vector Machine A Support Vector Machine (SVM)

It performs classification by finding the hyper plane that maximizes the margin between the classes. The vectors (cases) that define the hyper plane are the support vectors [19]
Algorithm
• Define an optimal hyper plane: maximize margin.
• Extend the above definition for non-linearly separable Problems: have a penalty term for
misclassifications
• Map data to high dimensional space where it is easier to classify with linear decision surfaces: reformulate problem so that data is mapped implicitly to this space. For this type of SVM, training involves the minimization of the error function.

(E) Logistic Regression

It is a classification algorithm ,used when the value the value of the target variable is categorical in nature.It is most commonly used when the data in question has binary output ,so when it belongs to one class or another or is either a 0 or 1,
Algorithm
• We apply Sigmoid Function,is a function that resembles an "S" shaped cuve when plotted on a graph.It takes values between 0 and 1 and "squishes" them towards the margins at the top and bottom,labeling them as 0 or 1.
• Equation for Sigmoid Function is:

y=1/(1+e^-x)

e -- represents the exponential function or exponential constant , value of approximatelly 2.71828
• This gives a value y that is extremely close to 0 if x is a large negative value and close to 1 if x is a large positive value. After the input value has been squeezed towards 0 or 1, the input can be run through a typical linear function, but the inputs can now be put into distinct categories

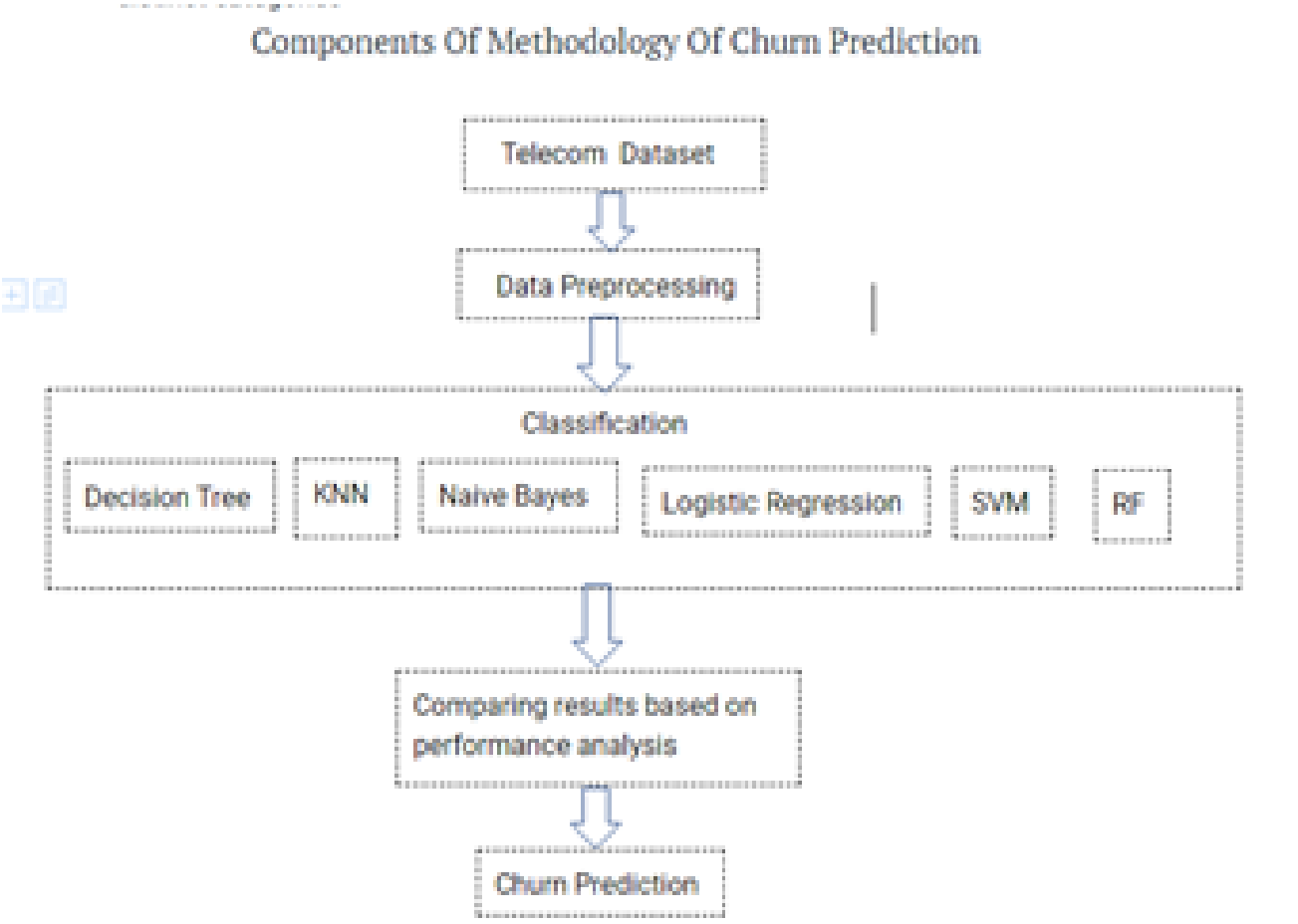
(F) Random Forest

The random forest algorithm is a supervised classification algorithm. As the name suggests, this algorithm creates the forest with a number of trees.
In general, the **more trees in the forest** the more robust the forest looks like. In the same way in the random forest classifier, the **higher the number** of trees in the forest gives **the high the accuracy** results.
If you know the decision tree algorithm. You might be thinking are we creating more number of decision trees and how can we create more number of decision trees. As all the calculation of nodes selection will be the same for the same dataset.
Yes. You are true. To model more number of decision trees to create the forest you are not going to use the same apache of constructing the decision with information gain or Gini index approach.
If you are not aware of the concepts of decision tree classifier, Please spend some time on the below articles, As you need to know how the decision tree classifier works before you learning the working nature of the random forest algorithm.

Random Forest pseudocode:

1. Randomly select **"k"** features from total **"m"** features.
    a. Where **k << m**

2. Among the **“k”** features, calculate the node **“d”** using the best split point.
3. Split the node into **daughter nodes** using the **best split**.
4. Repeat **1 to 3** steps until “l” number of nodes has been reached.
5. Build forest by repeating steps **1 to 4** for “n” number times to create **“n” number of trees**.



V. RESULTS AND DISCUSSION

The result of each classifier has been evaluated using different evaluation metrics and validated against overfitting using 10-fold cross-validation. The nested cross-validation approach also has been applied for the purpose of tuning the models' parameters. The experiments are conducted using Python 3.3 programming language through the Jupyter Notebook web application. Several libraries from Sciket-learn [35] have been used, which is a free software for the machine learning library in Python. The evaluation measures considered in this study are accuracy using F1-measure, sensitivity,specificity, and area under the curve (AUC).

This study is carried to predict weather the customer is leaving or not. This Prediction model is created in Python programming language. In our classification model we have used all classification algorithms; the classification algorithms were applied to the same data set.Dataset of 10,00 people with total of 14 service details like customer id,estimated salary,credit scores etc is used.

ACCURACIES OF ML MODELS

- Logistic Regression- 80.75%
- Desicion Tree- 77.1%
- Random Forest- 85.2%
- K-Nearest Neighbor- 83.2%
- Naive Bayes- 82.7%
- Support Vector Machine- 86.45

VI .ADVANTAGES AND DISADVANTAGES

Advantages:

- Access all the relevent data seamlessly and quickly.
- Segment the customers based on behaviour and demographics to improve retention.

● Deliver tailored promotions and offers to positively influence their behaviour.
● Minimize acquistion costs and increases marketing efficiency.
● Keep customers engaged and loyal.
● Predicting customers overall satisfaction as well as their experience with service quality.
● Identifying potential network issues, competitive threats, and at-risk customers.
● Identifying the negative customer experience trends and reducing attrition levels.
● Building a robust predictive model and gathering data.
● Creating new opportunities for cross-selling and upselling.

Disadvantages:

● A customer's lifetime value and the growth of the business maintains a direct relationship between each other i.e., more chances that the customer would churn, the less is the potential for the business to grow.
● Even a good marketing strategy would not save a business if it continues to lose customers at regular intervals due to other reasons and spend more money on acquiring new customers who are not guaranteed to be loyal.
● There is a lot of debate surrounding customer churn and acquiring new customers because the former is much more cost-effective and ensures business growth.
● Thus companies spend almost seven times more effort, and time to retain old customers than acquire a new one.
The global value of a customer lost is nearly two hundred, and forty-three dollars which makes churning a costly affair for any business.

VII. APPLICATIONS

● Telephone service companies
● Internet service providers
● Pay TV companies
● Insurance firms

- alarm monitoring services
- Predicting Insurance Customer Churn
- Lifetime Value

VIII. CONCLUSION AND FUTURE SCOPE

The main purpose of the application is to build a Machine Learning model to predict the customer churn.The model is built .We are developing a web application which is built using flask.Accurate prediction of customer churn is one of the emerging topics in telecommunication industry . Even though some approaches using real-time features shows very good performance in terms of accuracy. This work proposes a classification model to predict the customer churn using various machine learning algorithms.
All the five classification algorithms have been considered for customer churn.
From the above results, the objective is to find the better model for customer churn. The Support Vector Machine is the better model for predicting customer churn as it attains the accuracy of 86%.
Companies usually make a distinction between voluntary churn and involuntary churn. Voluntary churn occurs due to a decision by the customer to switch to another company or service provider, involuntary churn occurs due to circumstances such as a customer's relocatoion to a long-term care facility, death, or the relocation to a distant location . In most applications, involuntary reasons for churn are excluded from the analytical models. Analysts tend to concentrate on voluntary churn, because it typically occurs due to factors of the company-customer relationship which company's control, such as how billing interactions are handled or how after-sales help is provided. Predictive analytics use churn prediction models that predict customer churn by assessing their propensity of risk to churn. Since these models generate a small prioritized list of potential defectors, they are effective at focusing customer retention marketing programs on the subset of the customer base who are most vulnerable to churn

Telephone service companies, Internet service providers, pay TV companies, insurance firms, and alarm monitoring services, often use customer attrition analysis and customer attrition rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one. Companies from these sectors often have customer service branches which attempt to win back

defecting clients, because recovered long-term customers can be worth much more to a company than newly recruited clients. This can help companies in fetching information of active subscribers, collecting data and storing them

IX. REFERENCES

1. V. Lazarov and M. Capota, Churn prediction, Bus. Anal. Course, TUM Comput. Sci, Technische Univ. MÃ¼nchen, Tech. Rep., 2007. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.7201&rep=rep1&type=pdf.

2. Vadakattu, B. Panda, S. Narayan, and H. Godhia, Enterprise subscription churn prediction, in Proc. IEEE Int. Conf. Big Data, Nov. 2015, pp. 13171321.

3. S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman, Telecommunication subscribers churn prediction model using machine learning, in Proc. 8th Int. Conf. Digit. Inf. Manage., Sep. 2013, pp. 131136

4. A Novel Approach for Churn Prediction Using Deep Learning https://www.researchgate.net/publication/328819998_A_Novel_Appoach_for_Churn_Prediction_Using_Deep_Learning

5. S. A Survey on Customer Churn Prediction using Machine Learning Techniques.https://www.researchgate.net/publication/310757545_A_Survey_on_Customer_Churn_Prediction_using_Machine_Learning_Tech niques

6. Deep Learning in Customer Churn Prediction: Unsupervised Feature Learning on Abstract Company Indepenent Feature Vectors https://arxiv.org/pdf/1703.03869.pdf

7. Automated Feature Selection and Churn Prediction usingDeepLearning Models. https://www.irjet.net/archives/V4/i3/IRJET-V4I3422.pdf

8. Effectual Predicting Telecom Customer Churn using Deep Neural Network.https://www.ijeat.org/wp-content/uploads/papers/v8i5/D6745048419.pdf

9. Customer Churn Prediction in Telecommunication with Rotation Forest Methodhttps://www.researchgate.net/publication/282981765_Customer

10. _churn_prediction_in_telecommunication

11. Customer churn prediction in telecom using machinelearninghttps://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6