Prajwal Jagadish

Matriculation number: 31-04-067-20-2

Course of study: Msc in business intelligence and data science

Module: Business analytics

Topic: Machine learning and artificial intelligence

Abstract:

Machine learning is a class of algorithms that learn from examples, than explicitly performed in programming task. Machine learning algorithms formulate set of rules according to the example. This is the basis of artificial intelligence; it can be further divided into deep learning depending on the complexity and structure of the algorithm. Knime analytics platform is a open software to integrate and perform different data science models. Knime analytical platform is used to structure the models with two different applications applied in business. In methodology, the first model is based on customer churn in a telecommunication industry that describes about which customers are churning the contract. The other model is based on credit card fraud detection. The goals and objectives are the main features in analyzing in the whole process. The methodology describes about which machine learning algorithms are used with detailed benefits. The last part explains about the future works with respect to two different applications applied in business.

Introduction:

In developed countries, telecommunication sector is one of the important industries. In this competitive world industries are thriving hard to survive through implementing various strategies. Operators and technical progress have increased the level of competition in telecommunication industry. In customer churn, it is divided into two groups – (i) accidental churn: the conditions of a company keeps changing with regard to policies and standards, this creates a customer to adopt or change according to the utilization of services. For instance, the benefit of cost increases unreliably, that impacts customer to churn. (ii) Intentional churn: this happens when clients shift from one company to another that gives better services, cost-friendly, and ideas from the competitors. With regard to Telecommunication Company, the key factor is to generate revenues and up-sell the existing customers. The operators and technical members must identify clients before they exit from the company. However, establishing a unique classifier which predicts the future churn customers is essential in the whole process. There are various algorithms used in predicting customer churn, like k-nearest neighbor, decision tree, random forest, linear regression, neural networks and so on. The data of the customer churn is processed data and according to these data few well-known machine learning algorithms are applied. Therefore, selecting right attributes according to the data set and factors influencing churn customers is important to predict churned customers.

Goals: (benefits for a company)

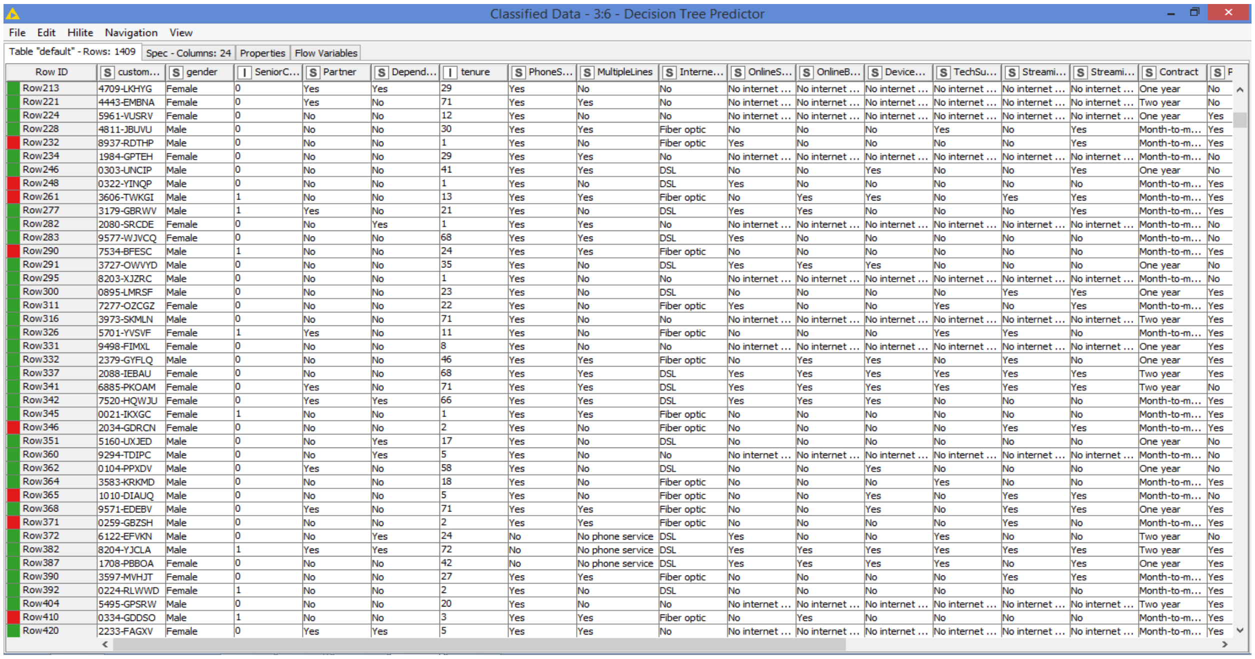
The objective from the problem statement is to build an effective machine learning model that predicts customer churn. The purpose of predicting is to evaluate various models with prescribed algorithms. Measuring performance of each model is essential which impacts the process of customer churn. This creates to give better insights and specific solutions for a company through influencing factors to avoid customer churn. On the whole process, by evaluating these factors telecommunication industry could implement better strategies.

Methodology (data with MLAIT used)

The data set is collected via web search of Reliance Company; the data set includes 7043 records with 21 variables. The data set contains following information:

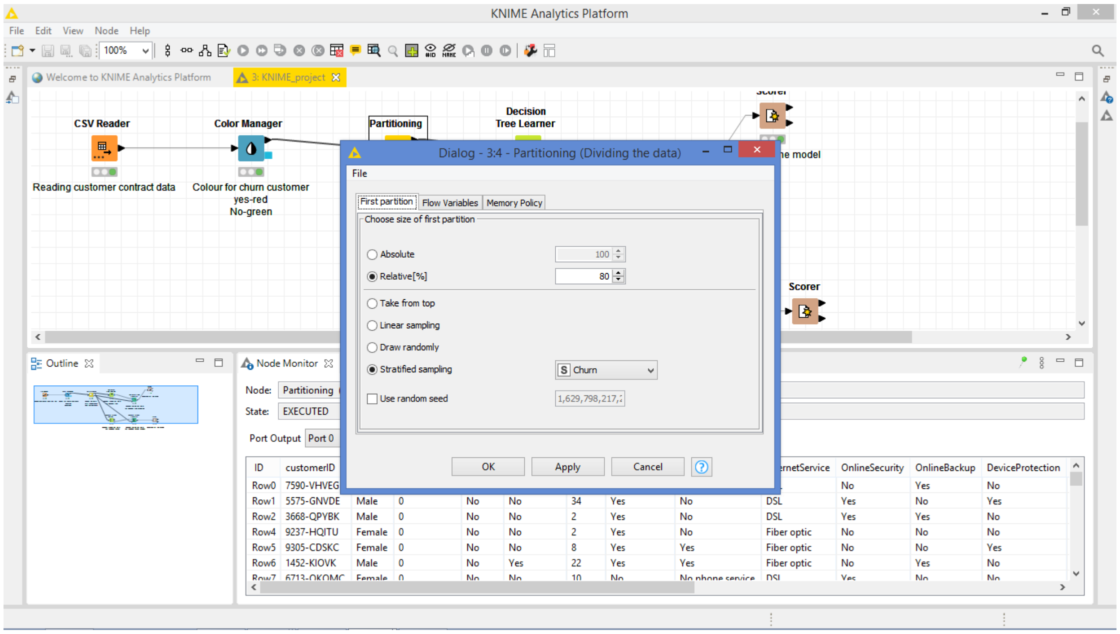
- Customers enrolled for different services and subscriptions: internet service, online security, online backup, streaming movies and TV, tech support and phone services.
- Exit of customers: the column name is Churn.
- Information about customers: Payment methods, monthly charges, total charges, paperless billing and Contract of the customer.
- Demographic information: Partners (dependents), gender, and senior citizen.

After analyzing the data set, the first step in Knime analytical software is to read the data from particular reader which the file belongs to. Here, the file is in Comma separated file (CSV). Therefore, CSV reader from the node repository is taken. Then, the CSV reader is executed.


Classified Data - 3:6 - Decision Tree Predictor
File
Edit
Hilite
Navigation
View
Table "default" - Rows: 1409
Spec - Columns: 24
Properties
Flow Variables

Row ID	S custom...	S gender	I SeniorC...	S Partner	S Depend...	I tenure	S PhoneS...	S MultipleLines	S Interne...	S OnlineS...	S OnlineB...	S Device...	S TechSu...	S Streami...	S Streami...	S Contract	S F
🟩 Row213	4709-LKHYG	Female	0	Yes	Yes	29	Yes	No	No	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	One year	No
🟩 Row221	4443-EMBNA	Female	0	Yes	No	71	Yes	Yes	No	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	Two year	No
🟩 Row224	5961-VUSRV	Female	0	No	No	12	Yes	No	No	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	One year	Yes
🟩 Row228	4811-JBUVU	Male	0	No	No	30	Yes	Yes	Fiber optic	No	No	No	Yes	No	Yes	Month-to-m...	Yes
🟥 Row232	8937-RDTHP	Male	0	No	No	1	Yes	No	Fiber optic	Yes	No	No	No	No	Yes	Month-to-m...	Yes
🟩 Row234	1984-GPTEH	Female	0	No	No	29	Yes	Yes	No	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	Month-to-m...	No
🟩 Row246	0303-UNCIP	Male	0	No	No	41	Yes	Yes	DSL	No	No	Yes	No	No	Yes	One year	No
🟥 Row248	0322-YINQP	Male	0	No	No	1	Yes	No	DSL	Yes	No	No	No	No	No	Month-to-m...	Yes
🟥 Row261	3606-TWKGI	Male	1	No	No	13	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to-m...	Yes
🟩 Row277	3179-GBRWV	Male	1	Yes	No	21	Yes	No	DSL	Yes	Yes	No	No	No	Yes	Month-to-m...	Yes
🟩 Row282	2080-SRCDE	Female	0	No	Yes	1	Yes	Yes	No	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	Month-to-m...	No
🟩 Row283	9577-WJVCQ	Female	0	No	No	68	Yes	Yes	DSL	Yes	No	No	No	No	No	Month-to-m...	No
🟥 Row290	7534-BFESC	Male	1	No	No	24	Yes	Yes	Fiber optic	No	No	No	No	No	No	Month-to-m...	Yes
🟩 Row291	3727-OWVYD	Male	0	No	No	35	Yes	No	DSL	Yes	Yes	Yes	No	No	No	One year	No
🟩 Row295	8203-XJZRC	Male	0	No	No	1	Yes	No	No	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	Month-to-m...	No
🟩 Row300	0895-LMRSF	Male	0	No	No	23	Yes	No	DSL	No	No	No	No	Yes	Yes	One year	Yes
🟩 Row311	7277-OZCGZ	Female	0	No	No	22	Yes	No	Fiber optic	Yes	No	No	Yes	No	Yes	Month-to-m...	Yes
🟩 Row316	3973-SKMLN	Male	0	No	No	71	Yes	No	No	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	Two year	Yes
🟩 Row326	5701-YVSVF	Female	1	Yes	No	11	Yes	No	Fiber optic	No	No	No	Yes	Yes	No	Month-to-m...	Yes
🟩 Row331	9498-FIMXL	Female	0	No	No	8	Yes	No	No	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	One year	Yes
🟩 Row332	2379-GYFLQ	Male	0	No	No	46	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	No	One year	Yes
🟩 Row337	2088-IEBAU	Female	0	No	No	68	Yes	Yes	DSL	Yes	Yes	Yes	Yes	Yes	Yes	Two year	Yes
🟩 Row341	6885-PKOAM	Female	0	Yes	No	71	Yes	Yes	DSL	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No
🟩 Row342	7520-HQWJU	Female	0	Yes	Yes	66	Yes	Yes	DSL	Yes	Yes	Yes	No	No	No	Month-to-m...	Yes
🟩 Row345	0021-IKXGC	Female	1	No	No	1	Yes	Yes	Fiber optic	No	No	No	No	No	No	Month-to-m...	Yes
🟥 Row346	2034-GDRCN	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	Yes	Yes	Month-to-m...	Yes
🟩 Row351	5160-UXJED	Male	0	No	Yes	17	Yes	No	DSL	No	No	No	No	No	No	One year	No
🟩 Row360	9294-TDIPC	Male	0	No	Yes	5	Yes	No	No	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	Month-to-m...	No
🟩 Row362	0104-PPXDV	Male	0	Yes	No	58	Yes	No	DSL	No	No	Yes	No	No	No	One year	No
🟩 Row364	3583-KRKMD	Male	0	No	No	18	Yes	No	Fiber optic	No	No	No	Yes	No	No	Month-to-m...	Yes
🟥 Row365	1010-DIAUQ	Male	0	No	No	5	Yes	No	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-m...	No
🟩 Row368	9571-EDEBV	Male	0	Yes	No	71	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	One year	Yes
🟥 Row371	0259-GBZSH	Male	0	No	No	2	Yes	Yes	Fiber optic	No	No	No	No	Yes	No	Month-to-m...	Yes
🟩 Row372	6122-EFVKN	Male	0	No	Yes	24	No	No phone service	DSL	Yes	No	No	Yes	No	No	Two year	No
🟩 Row382	8204-YJCLA	Male	1	Yes	Yes	72	No	No phone service	DSL	Yes	Yes	Yes	Yes	Yes	Yes	Two year	Yes
🟩 Row387	1708-PBBOA	Female	0	No	No	42	No	No phone service	DSL	Yes	Yes	Yes	Yes	No	Yes	One year	Yes
🟩 Row390	3597-MVHJT	Female	0	No	No	27	Yes	Yes	Fiber optic	No	No	No	No	Yes	Yes	Month-to-m...	Yes
🟩 Row392	0224-RLWWD	Female	1	No	No	2	Yes	No	DSL	No	No	No	No	No	No	Month-to-m...	Yes
🟩 Row404	5495-GPSRW	Male	0	No	No	20	Yes	No	No	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	Two year	Yes
🟥 Row410	0334-GDDSO	Male	1	No	No	3	Yes	Yes	Fiber optic	No	Yes	No	No	No	No	Month-to-m...	Yes
🟩 Row420	2233-FAGXV	Female	0	Yes	Yes	5	Yes	Yes	No	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	No internet ...	Month-to-m...	Yes

Each row have customer ID which is the primary key to associate the data set. Knime analytical software performs various workflows and nodes; one of them is color manager that helps us to analyze each row colored along with customer-ID. Here, the color manager is used to analyze whether the customer is churning or not. By this it will be easier to explain and understand each customer's transaction. If the customer is churned (Yes) it is colored with red and not churned (No) with Green.



Dialog - 3:4 - Partitioning (Dividing the data)

File

First partition | Flow Variables | Memory Policy

Choose size of first partition

○ Absolute 100

● Relative[%] 80

○ Take from top

○ Linear sampling

○ Draw randomly

● Stratified sampling S Churn

□ Use random seed 1,629,798,217,2

OK Apply Cancel

Node: Partitioning

State: EXECUTED

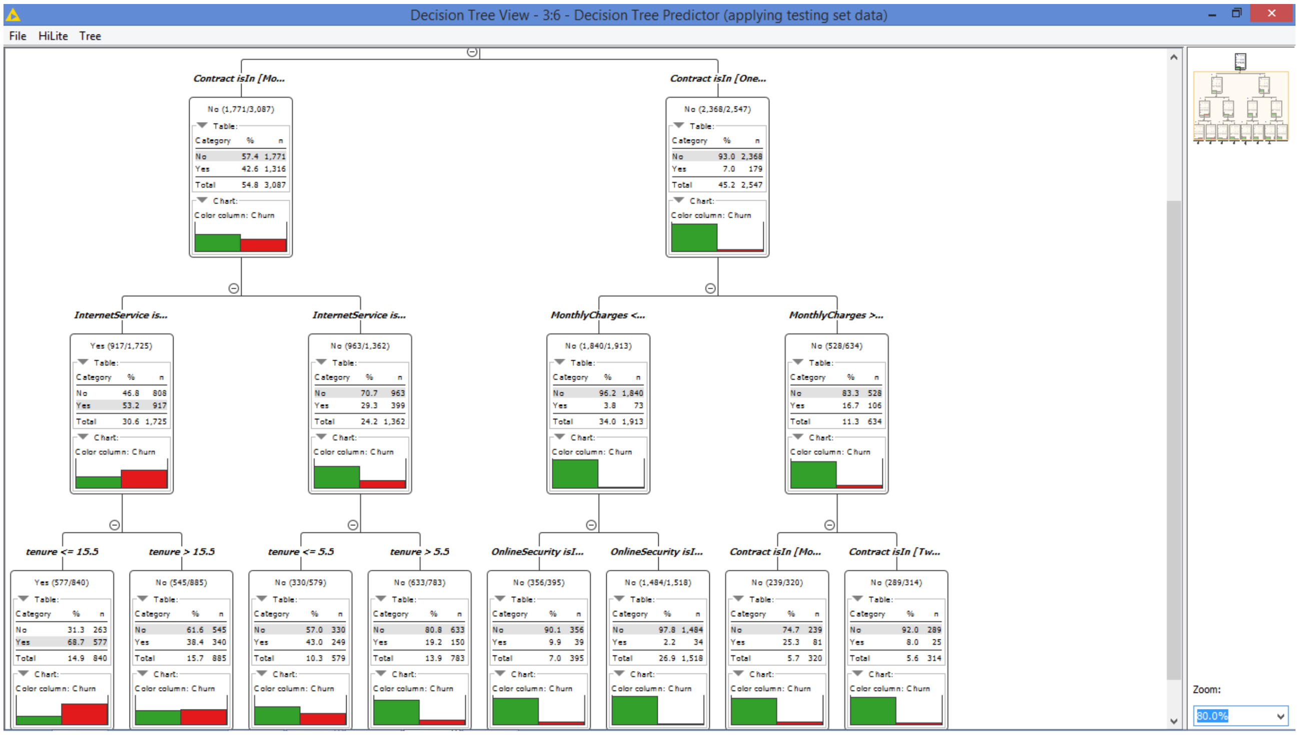
Port Output Port 0

ID	customerID								ernetService	OnlineSecurity	OnlineBackup	DeviceProtection
Row0	7590-VHVEG									No	Yes	No
Row1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes
Row2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No
Row4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No
Row5	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes
Row6	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No
Row7	6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes	No	No

There are three stages which are generally characterized in customer churn, which are training phase, testing phase, and prediction phase. In this process, there is a partition node which is applied to divide the data set into two parts that is training set and testing set. In this phase, partition node is taken and executed by configuring relative percentage which directs to the training set. Here, the relative percentage is 80%, which means assigning 80% of the entire data set into training part, and remaining 20% into the test data set. There are multiple methods which are performed for data selection like linear sampling, draw randomly, and stratified sampling. Here, with respect to customer churn, stratified sampling is used for the selection in the data set, because stratified sampling divides the population into smaller groups according to the features available in the data set. This is essential in selecting the data according to the data set. However, after selecting these in the configuration window the nodes are connected to the predicting stage.

After analyzing and assigning the data into partition node, the next step of the process is to choose which machine learning algorithm is suitable according to the data set. As discussed earlier there many machine learning algorithms that can evaluate the performance of the data set. Choosing the right model in this process is a challenging move. By evaluating the features of the data set and trying different models accordingly is essential. With regard to customer churn, the prediction is dependent on feature which is influencing the customers to churn. Many algorithms like SVM, ANN, decision tree, random forest have been reviewed but these models depend on the accuracy.

Decision tree and random forest are implemented in this process. According to the customer churn data; there were many groups which influences the customer to churn. So, these algorithms help in grouping the data with respect to the features.
The partition node of training set are connected to decision tree learner, in this node the mode is trained. Then the other node is connected to decision tree predictor node where the test set data are attached to give prescribed predictions.



Decision Tree View - 3:6 - Decision Tree Predictor (applying testing set data)

Contract isIn [Mo...

No (1,771/3,087)

Category	%	n
No	57.4	1,771
Yes	42.6	1,316
Total	54.8	3,087

Chart: Color column: Churn

Contract isIn [One...

No (2,368/2,547)

Category	%	n
No	93.0	2,368
Yes	7.0	179
Total	45.2	2,547

Chart: Color column: Churn

InternetService is...

Yes (917/1,725)

Category	%	n
No	46.8	808
Yes	53.2	917
Total	30.6	1,725

Chart: Color column: Churn

InternetService is...

No (963/1,362)

Category	%	n
No	70.7	963
Yes	29.3	399
Total	24.2	1,362

Chart: Color column: Churn

MonthlyCharges <...

No (1,840/1,913)

Category	%	n
No	96.2	1,840
Yes	3.8	73
Total	34.0	1,913

Chart: Color column: Churn

MonthlyCharges >...

No (528/634)

Category	%	n
No	83.3	528
Yes	16.7	106
Total	11.3	634

Chart: Color column: Churn

tenure <= 15.5

Yes (577/840)

Category	%	n
No	31.3	263
Yes	68.7	577
Total	14.9	840

Chart: Color column: Churn

tenure > 15.5

No (545/885)

Category	%	n
No	61.6	545
Yes	38.4	340
Total	15.7	885

Chart: Color column: Churn

tenure <= 5.5

No (330/579)

Category	%	n
No	57.0	330
Yes	43.0	249
Total	10.3	579

Chart: Color column: Churn

tenure > 5.5

No (633/783)

Category	%	n
No	80.8	633
Yes	19.2	150
Total	13.9	783

Chart: Color column: Churn

OnlineSecurity isI...

No (356/395)

Category	%	n
No	90.1	356
Yes	9.9	39
Total	7.0	395

Chart: Color column: Churn

OnlineSecurity isI...

No (1,484/1,518)

Category	%	n
No	97.8	1,484
Yes	2.2	34
Total	26.9	1,518

Chart: Color column: Churn

Contract isIn [Mo...

No (239/320)

Category	%	n
No	74.7	239
Yes	25.3	81
Total	5.7	320

Chart: Color column: Churn

Contract isIn [Tw...

No (289/314)

Category	%	n
No	92.0	289
Yes	8.0	25
Total	5.6	314

Chart: Color column: Churn

Zoom: 80.0%

The above diagram explains about decision tree model, where different tree are created according to the features influencing the customers to churn. In the first tree it explains if the customers had monthly contract are more likely to churn than the yearly contract. With regard to internet service 53% of the people are likely to churn. With both parameter of internet service and monthly charges, 63% of the people are churning the contract. Less the tenure, the churn of the contract increases. On the other side, people having online security, Long term tenure, and Streaming movies and TV are most likely that they don't churn the contract. However, decision tree model's accuracy is 75%. That means this model holds good in predicting customer churn data set.

Random forest predictor node is also connected as the same process like decision tree. In this model the prediction of churn customer is quite accurate. Therefore the accuracy of this model is 76% that is greater than decision tree by 1%.

Conclusion:

From the predictions made by the machine learning model can help in understanding the customers, who might churn or exit from their services. The company can utilize this information to implement better strategies. The following factors can help the company in many different ways:

- Middle level management and Relationship manager will get effective information about influencing factors of churned customer.
- This can help and fix their problems by discussing customer needs and pain points before they churn.
- By implementing these strategies can reduce operational cost on call centers
- This in turn will help in increasing company value and the retention of their customer.

Additional issues that can be discussed further are:

- The company can collect more information from customers like reviews, ratings, and so on, to understand the patterns which lead for attrition.
- These models can be evaluated and deployed on a daily basis so that the changes of the customer can act accordingly and immediately.
- Getting data from other telecommunication industries will help in analyzing the different strategies and market share.
- Finding out and getting information about geographic information about the customer helps in analyzing patterns around customer churn with and geography few demographic attributes.

The next different application used in machine learning is Credit card Fraud detection:

Credit card fraud detection is a well-known aspect but a tricky problem to solve. Initially, evaluating the pattern from limited data is a challenging process to detect the fraudulence of credit cards. Secondly, the data set is very large with many transactions and entries from fraudsters which lead to many constraints. Data sets are not available for the general public and the reviews from the researchers are hidden. Lastly, the changing profiles and continuous advancements of fraudulent behaviors have made the process different. For instance, the fraudulent transaction may have occurred in the present or vice versa, so the transaction behavior is always different.

With respect to machine learning models, different techniques are used to analyze different measures of fraud actions. Features are constructed according to the sample fraudulent data sets. To detect credit card frauds, behavioral patterns are compared with past history of transactions and features such as daily expenses, location, and time of the transaction. Therefore, by evaluating all the above patterns, different models can be built on this basis.

Problem definition:

Online payment does not need a physical card; by this, anyone can access it and is aware of the card details that lead to fraud transactions. When fraud transactions occur the cardholder will get a notice about the transaction. However, to avoid fraud transactions various machine learning algorithms are applied.

Methodology: (Data transformation and MLAIT)

To detect the fraud KNN (l-nearest neighbor) method is used. The agenda of proposed system is to create credit Card Fraud Detection System Using Machine Learning.

The data set is taken from Kaggle, which consist of transactions related to simulated mobile based payment. The data set contains 496 frauds out of 280000 transactions. Dataset contain variables of numerical input which are made by PCA transformations. With data confidentiality issues background, information and original features are not given. Different variables like V1, V2,…. and so on are created with principal component analysis. Time and Amount are the two features which are not transformed from PCA, In which Amount consist of all the price in the transaction and Time consists of seconds elapsed between each transaction with first transaction of the data set. In Class attribute takes value 1 (fraud) and 0 otherwise.

KNN model have been used to detect several techniques. According to reviews and applied algorithms K-nearest neighbour stood out to be the best model in Supervised learning, where the query is based on majority of KNN model. The three main factors which influence to choose KNN algorithm is:

- To locate K-nearest neigbhour distance metric is used
- To derive a classification in KNN distance rule is used.
- The new sample is classified by number of neighbor.

Therefore, by analyzing these parameters with various methods of supervised learning, k-nearest neighbor have high performance in detecting fraud.

The classification of any incoming transaction with nearest point of other new transactions is made to analyze fraud. If any nearest neighbor occurs as fraudulence, then the indication of transaction is fraud.

The noisy data can be reduced with the help of larger k-values. The distance between two data can be calculated in various ways. Euclidean distance is a better choice for continuous variables. By optimizing the distance metric the performance of KNN can be improved. By this feature selection is being taken to perform training set and testing set. The model building is verified and the accuracy is being tested through validating the test set. However, the performance is evaluated based on precision, sensitivity and accuracy analysis.

Conclusion and future work:

With imbalanced data set, Out of 3200 points there were 12 data points which belonged to Class 1 that is under fraudulence. The remaining of the dataset is class 0. Therefore, the model of k-nearest neigbhour has performed well according to each transaction. This can be implemented in all the companies because every company uses net banking facilities, in which the payment is done by online transaction.

This becomes difficult when the policies are restricted and cannot investigate any transaction without any legal documents. This process can be done in most of the companies as every company is digitalized today. Fraud detection play important role to minimize crime activities, this can be even extracted with respect to the policies.