**Accusaga Data science Assignment**

**Customer retention Model**

Submitted by:

Soham Bandgar

# ACKNOWLEDGMENT

The different references that helped me and guided me in completion of the project are as follows:-

1. **Customers Churn Prediction in Financial Institution Using Artificial Neural Network.**

Kamorudeen A. AMUDA, Adesesan B. ADEYEMO Published in Department of Computer Science, University of Ibadan, Nigeria.

1. **Practical Data Mining: Applications, Experiences and Challenges** and Organized by: Markus Ackermann, Carlos Soares, Bettina Guidemann; In Partnership with SAS Deutschland, Heidelberg.
2. Bhattacharya, C. B. (1994). **When Customers Are Members: Customer Retention in Paid Membership Contexts**
3. **Knowledge Based Customer Churan Prediction For Telecom Services Using Python** Computer Science & Engineering under Maulana Abul Kalam Azad University of

Technology by **Pranjal Chowdhury**

# INTRODUCTION

## Business Problem Framing:-

* Customer satisfaction has emerged as one of the most important factors that guarantee the success of online store; it has been posited as a key stimulant of purchase, repurchase intentions and customer loyalty.
* A comprehensive review of the literature, theories and models have been carried out to propose the models for customer activation and customer retention. Five major factors that contributed to the success of an e-commerce store have been identified as: service quality, system quality, information quality, trust and net benefit.

### Conceptual Background of the Domain Problem:-

**Points to Remember:**

There is no missing values with dataset with the aid of Python programming language libraries. To achieve better performance, the categorical data was transformed to numerical format using the Label Encoder function in Python. Feature scaling was applied to normalize the data and improved the computational time.

* There are no null values in the dataset.
* The dataset is balanced. Label ‘1’ has approximately 50% records, while, label ‘0’ has approximately 50% records.
* For some features, there may be values which might not be realistic. You may have to observe them and treat them with a suitable explanation.
* You might come across outliers in some features which you need to handle as per your understanding. Keep in mind that data is expensive and we cannot lose more than 78% of the data.

### Review of Literature:-

Data mining is the process of analyzing data from different perspectives and extracting useful knowledge from it. It is the core of knowledge discovery process. The various steps involved in extracting knowledge from raw data as depicted in figure-1. Different data mining techniques include classification, clustering, association rule mining, prediction and sequential patterns, neural networks, regression etc. Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to classification technique. This approach frequently employs Decision tree based classification Algorithm. In classification, a training set is used to build the model as the classifier which can classify the data items into its appropriate classes. A test set is used to validate the model.



Figure -1

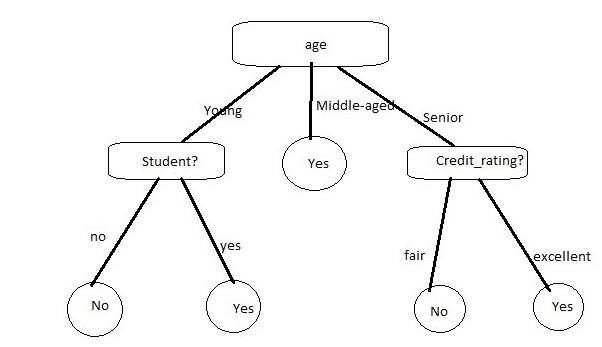
#### Data Mining in Banking

Due to tremendous growth in data the banking industry deals with, analysis and transformation of the data into useful knowledge has become a task beyond human ability. Data mining techniques can be adopted in solving business problems by finding patterns, associations and correlations which are hidden in the business information stored in the database. By using data mining techniques to analyse patterns and trends, bank executives can predict, with increased accuracy, how customers will react to adjustments in interest rates, which customers are likely to accept new product offers, which customers will be at a higher risk for defaulting on a loan, and how to make customer relationships more profitable. Globalization and the stiff competition had led the banks focus towards customer retention and fraud prevention. To help them for the same, data mining is used. By analyzing the past data, data mining can help banks to predict credible customers. Thus they can prevent frauds, they can also plan for launching different special offers to retain those customers who are credible. Certain areas that effectively utilize data mining in banking industry are marketing, risk management and customer relationship management.

#### Proposed Model

**Machine learning: Decision Tree:-**

Decision tree algorithm in machine learning methods which efficiently performs both classification and regression tasks. It creates decision trees. Decision trees are widely used in the banking industry due to their high accuracy and ability to formulate a statistical model in plain language. In Decision tree each node represents a feature (attribute), each link (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continues value).



### Motivation for the Problem Undertaken:-

Predicting churn, i.e. if a customer is about to leave for a competitor, is an important application of analysing customer behaviour. It is typically much more expensive to acquire new customers then to retain existing ones. In the telecommunication industry, for example, this factor is in the range of about five to eight [1]. Correctly predicting that a customer is going to churn and then successfully convincing him to stay can substantially increase the revenue of a company, even if a churn prediction model produces a certain number of false positives.

Beside the prediction of churn, other customer-related events like faults, purchases or complaints can be predicted in order to be able to resolve some problems before the actual event occurs. The prediction of sales events can be used for cross-selling, where a certain product is offered just to customers who have an increased likelihood to buy it.

An overview of the steps we'll take in this project are:-

1. Importing the libraries
2. Loading the dataset
3. Selecting relevant features
4. Converting categorical columns to numeric ones
5. Preprocessing the data
6. Training a machine learning algorithm
7. Evaluating the machine learning algorithm
8. Evaluating the dataset features

## Analytical Problem Framing:-

### Hardware and Software Requirements and Tools Used**:-**

**Anaconda:** is an open source software distribution of R and Python programming languages that are used for scientific computing such as data science, predictive analytics, machine learning, and deep learning applications purposely to simplify package management and deployment.

**Jupyter Notebook**: is an open free source web application that is used for data cleaning and transformation, numerical, simulation, statistical modelling, data visualization and so on. **Matplotlib:** is an amazing visualization library in Python programming language for twodimensional plots of arrays. One of the greatest benefits of visualization is that it allows high dimensionality data to visualize and easily understandable and it consists of several plots like line, bar, scatter, histogram and so on.

**Pandas:** is the most popular Python programming language package that offers powerful, expensive and flexible data structures that make data manipulation and analysis easy.

**Numpy:** is the fundamental package for scientific computing in Python programming language that contains a powerful N-dimensional array object and also useful in linear algebra.

**Label Encoder:** is a Python programming language package that is used to transform nonnumerical labels (or nominal categorical variables) to numerical labels.

**ANN visualizer**: is a Python programming language library that enables visualization of an artificial neural network.

**Seaborn:** is a Python data visualization library based on matplotlib that provides high-level interface for drawing attractive and informative statistical graphics. **Sci-kit Learn**: it is a free machine learning library for Python programming language that designed to interoperate with Python numerical Numpy and scientific libraries SciPy. Also, it can be used for classification, regression and clustering algorithms including support vector machine, linear and logistic regression, random forests, gradient boosting, decision tree, K-means and so on. **TensorFlow**: is an open source Python library used for machine learning applications such as neural network and used Keras as a backend.

**Keras**: is a neural network framework for Python programming language that provides a convenient way to define and train almost any kind of deep learning model.

**Neuro Solutions Infinity:** is one of the most powerful neural network software of the Neuro Solutions family that streamlines the data mining process by automatically cleaning and preprocessing data. It uses distributed computing, advanced neural networks and artificial intelligence (AI) to model data; creates highly accurate predictive models with an easy-to-use and intuitive interface that provides valuable insights that can be used to drive better decisions.

**Microsoft SQL Server:** is a relational database management system developed by Microsoft Incorporation which primarily used for storing and retrieving data requested by other software applications that run on the same computer or on another computer across a network.

**Flowchart**: illustrate the step by step procedure on how to predict customers churn with the multilayer perceptron of artificial neural network architecture.

### Data Inputs- Logic- Output Relationships:-

Customer\_ID- corresponds to the record (row) number and has no effect on the output

Var1-this is continuous data varies from 0 to 3000.

Var2 -This is numerical data varies from 0 to 300.

Group - This colum contain eight different types of groups .categorical in format .

Category- this column contain 5 different types of Category,categorical in format.

Rating- 3 different types of rating contain in categorical in format

**Purchased\_ABC\_product- this is our target column .which is balance in nature.**

### **Data Preprocessing Done:-**

The implementation was carried out in two phases; MLP software development on Python and then using Neuro Solution Infinity software. Multilayer Perceptron: MLP is an architecture of the artificial neural network that consists of multiple layers where each layer is fully connected with the next layer in a feed-forward direction. The first layer and the last layers represent the inputs (independent variables) and outputs (target variables) of the system respectively. Connections between the nodes represented as weights. The more the numbers of hidden layers in the network, the more the complexity of the network.

**Steps for MLP Model:-**

1. Extraction of customer data database.
2. Data pre-processing for removal of noise and transformation of categorical data into numerical data.
3. Data splitting (train and test set) and feature scaling
4. Start at the input layer by forwarding propagates patterns of the train data through the network to generate output.
5. Using cost function to calculate the network output to minimize the error rate.
6. Find the derivative with respect to each weight in the network and updating the model,
7. Calculate the network output and apply threshold function to obtain the predicted class label.
8. Evaluate the model.

## Model/s Development and Evaluation:-

### **Machine Learning Algorithm Evaluation:-**

Now that the algorithm has been trained, it's time to see how well it performs. For evaluating the performance of a classification algorithm, the most commonly used metrics are the [F1 MEASURE, PRECISION, RECALL and ACCURACY.](https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9) In Python's scikit-learn library, you can use built-in functions to find all of these values. Execute the following script:

Model 4 - Ensemble method :GradientBoostingClassifier

Hyperparameter Tuning:GRidSearch CV,Model Development,Cross validation

parameters={'loss':['deviance','exponential'],'criterion':['friedman\_mse','squared\_error']}

gd=GradientBoostingClassifier()

clf=GridSearchCV(gd,parameters)

clf.fit(x\_train,y\_train)

print(clf.best\_params\_)

best\_result=clf.best\_score\_

print(best\_result)

**{'criterion': 'friedman\_mse', 'loss': 'exponential'}**

**0.767586716337939**

**gd=GradientBoostingClassifier(criterion='friedman\_mse',loss='deviance')**

gd.fit(x\_train,y\_train)

gd.score(x\_train,y\_train)

predgd=gd.predict(x\_test)

print(accuracy\_score(y\_test,predgd))

print(confusion\_matrix(y\_test,predgd))

print(classification\_report(y\_test,predgd))

**0.7648394421018488**

**[[2173 835]**

**[ 615 2543]]**

**precision recall f1-score support**

**0 0.78 0.72 0.75 3008**

**1 0.75 0.81 0.78 3158**

**accuracy 0.76 6166**

**macro avg 0.77 0.76 0.76 6166**

**weighted avg 0.77 0.76 0.76 6166**

from sklearn.model\_selection import KFold

score=cross\_val\_score(gd,x,y,cv=5)

print(score)

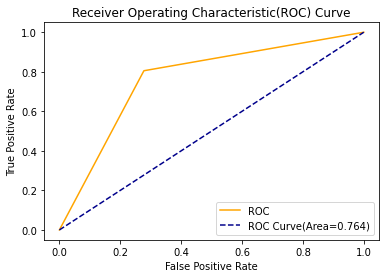
print(score.mean())

print(score.std())

**[0.76256893 0.77651638 0.76451508 0.76273111 0.76999189]**

**0.7672646769993415**

**0.005351547597010499**

****

The results indicate an accuracy of 75%, which means that our algorithm successfully predicts customer retention 76% of the time.

# CONCLUSION

The results indicate an accuracy of **76%,** which means that our algorithm successfully predicts customer retention 76% of the time. Customer churn prediction is crucial to the long-term financial stability of a company.

This marks the end of our process.

we have successfully trained our model to predict the customer data from Sample Data Sets with the aim of building and

comparing several customer churn prediction models with an accuracy of ~76%

sns.pairplot(dfs)

