Project Summary

Batch details	PGPDSE-BLR Oct22
Team members	Venkata Sai Pavan Teja Angina Akhil Anilkumar Shivaprasad G Sahil Kumar Meher Arun Sv
Domain of Project	Retail
Proposed project title	Telecom Customer Churn
Group Number	9
Team Leader	Sahil Kumar Meher
Mentor Name	Mrs.Pranita Mahajan

Date: 19 February 2023

Sahil Kumar Meher

Signature of the Mentor

Signature of the Team Leader

Table of Contents

Sl. No.	Topic	Page No
1	Overview	3
2	Business problem goals	3
3	Topic survey in brief	6
4	Critical assessment of topic survey	10
5	Methodology to be followed	10
6	References	14

Overview

The telecommunications sector has become one of the main industries in developed countries. The technical progress and the increasing number of operators have raised the level of competition. Companies are working hard to survive in this competitive market depending on multiple strategies.

Customer churn is a considerable concern in service sectors with highly competitive services. Onthe other hand, predicting the customers who are likely to leave the company will represent potentially large additional revenue source if it is done in the early phase.

Business problem statement

1. Business Problem Understanding

Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenue of the companies, especially in the telecom field, companies are seeking to develop means to predict potential customer to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn.

2. Business Objective

The main contribution of our work is to develop a churn prediction model which assists telecom operators to predict customers who are most likely to subject to churn.

3. Scope of the project

The scope of our project is comparative analysis of different ML models and classification of customers and predict whoare likely to churn from the available data and recommending the services to each customers so as to retain them and development of revenue model.

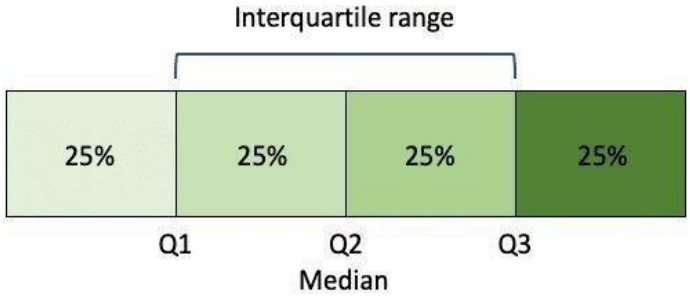
4. Approach

Data analysis, cleaning/pre-processing: The pre-processing of the dataset before performing ML functions involves the following:

a. **Descriptive Analysis:** Descriptive Analysis is used to describe the basic features of the data in a study. It provides summary of the data and the type of columns. Measures of variability help communicate the spread of distribution by describing the shape and spread of the data set.

b. **Treatment of Missing Values**: Detecting and handling missing values is important, as they can impact the results of the analysis, and there are algorithms that can't handle them. There are cases when a variable has a lot of missing values. In that case, we can drop the variable. For the categorical variable, the missing values can be replaced by the most frequent class of the variable. For the numeric variable, missing values can be replaced by the mean/ median.

c. **Treating Outliers:** Outlier is an observation in the data that lies at an abnormal distance from other values. Presence of Outliers, may result in inaccurate prediction. Hence it is necessary to remove them. The interquartile range is one of the measures of outlier treatment. It is the difference between the third quartile and the first quartile. The IQR gives the range ofmiddle 50% of the data which is free from outliers.



IQR = Q3 - Q1

d. **Encoding Categorical Variables**: Since, machine learning models are based on Mathematical equations and we can intuitively understand that it would cause some problem if we can either keep the Categorical data by encoding the categorical variable or we can drop by checking whether we need the variable for further modelling process because we would only want numbers in the equations.

e. **Dropping Unnecessary Columns**: We are removing the columns which do not contribute to the model building or the columns which are of less, or of no importance

f. **Removal / Replacing of Special Characters (if any):** Special characters such as '?', '$', '%' should be replaced with Nan so that they are easier to treat and replace or to remove.

g. **Scaling**: It helps to normalize the data within a particular range and as well as in speeding up the calculations in an algorithm.

Topic Survey in brief

1. Problem understanding

The reason behind describing customer churn in the preceding paragraphs is because, the goal of our next Machine Learning project is to develop an algorithm that can accurately predict customers who are most likely to churn. With the help of various visualization libraries that are at our disposal, we will be able to figure out possible parameters that govern a customer's decision to churn.

2. Current solution to the problem

Customer retention is extremely critical to the health of any business, regardless of size or industry. It is a common representation of a business's ability to keep its existing customers and maximize its revenue. It is important to dig into the factors that drive your customer retention rate and generate opportunities to improve your customer success strategy.

Current solution for customer churn:

- Competitive Pricing.
- Classification of customers.
- Service recommendation to each customers.
- Offer incentives.
- Frequent Feedback from Customers.
- Seamless customer service.
- Wide network coverage.
- Exclusive benefits for existing customers.
- Analyze churn when it happens.

3. Proposed solution to the problem

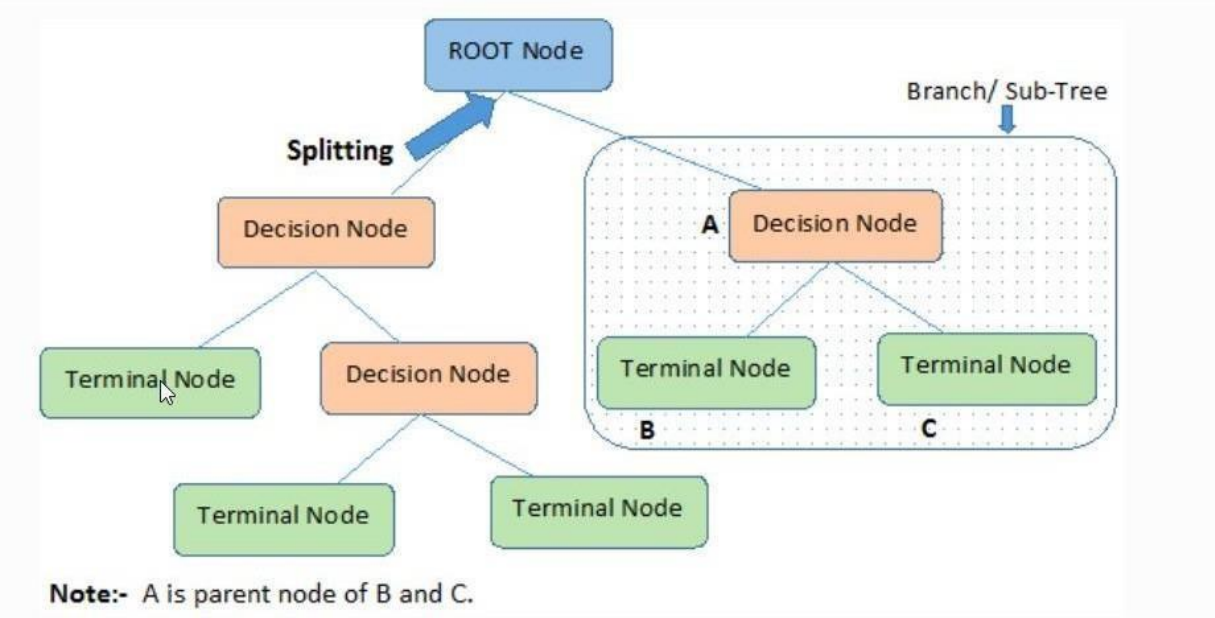
Exploratory Data analysis, Data Visualization, Building ML Models using different Algorithms to drive predictive analysis.

- **Decision Tree Algorithm:**

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

Types of Decision Trees

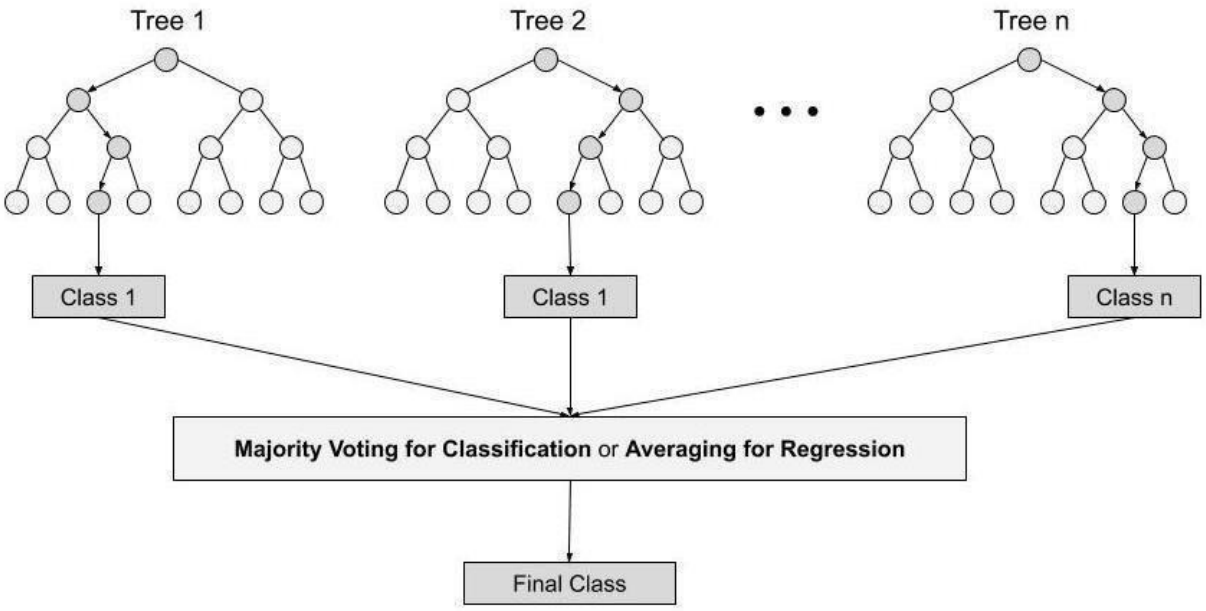
1. Categorical Variable Decision Tree
2. Continuous Variable Decision Tree



- **Random Forest**

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.



- **Logistic Regression**

Logistic regression is a transformation of a linear regression using the sigmoid function. The vertical axis stands for the probability for a given classification and the horizontal axis is the value of x. It assumes that the distribution of Bernoulli distribution. The formula of LR is as follows:

Here $\beta\_0+\beta\_1\ x$ is similar to the linear model y = ax + b. The logistic function applies a sigmoid

$$F(x)=1/\ (1+\ e\text{^}\ (-(\beta\_0+\beta\_1\ x)))$$

function to restrict the y value from a large scale to within the range 0–1.

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. As aspects of Retail and Marketing classification problems -Customer Churn, such logistic regression is a useful analytic technique.

Predictive models built using this approach can make a positive difference in your business or organization. Because these models help you understand relationships and predict outcomes, you can act to improve decision-making.

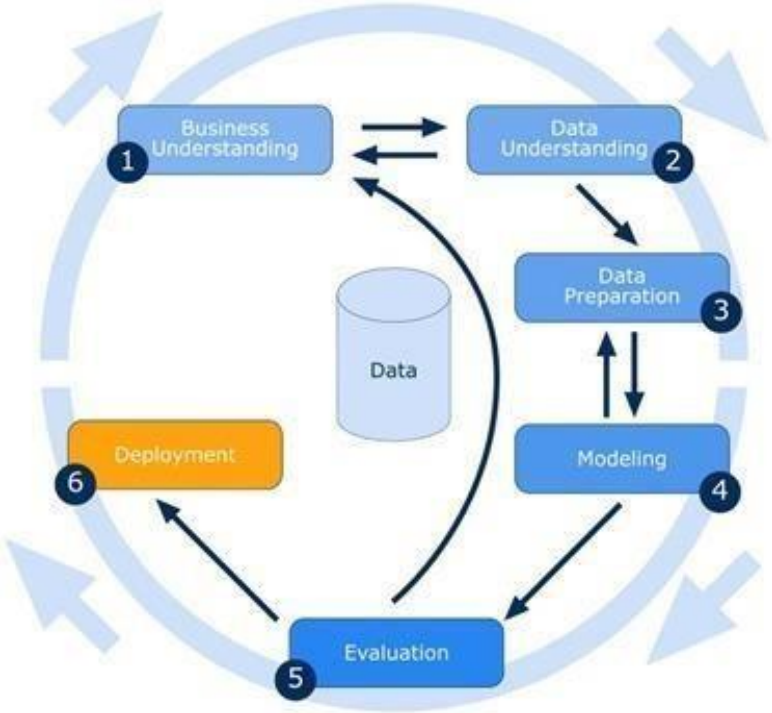
- **Gradient Boosting**

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that the errors in machine learning algorithms are broadly classified into two categories i.e. Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it is used to minimize bias error of the model.
Unlike, Ada boosting algorithm, the base estimator in the gradient boosting algorithm cannot be mentioned by us. The base estimator for the Gradient Boost algorithm is fixed and i.e. Decision Stump. Like, AdaBoost, we can tune the n_estimator of the gradient boosting algorithm. However, if we do not mention the value of n_estimator, the default value of n_estimator for this algorithmis 100.
Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier). When itis used as a regressor, the cost function is Mean Square Error (MSE) and when it is used as a classifier then the cost function is Log loss. We want our predictions, such that our loss function (MSE) is minimum.

Critical Assessment of Topic Survey

Once the customers at risk of churning have been identified, the **customer retention team** has to know exactly what marketing action to run. The reasons for churning out differs from person to person. Hence, it is critical to practice 'targeted proactive retention'. This means knowing in advance which marketing action will be the most effective for each and every customer.

Methodology to be followed



- **Business Understanding:**

It's all about understanding the overview, the aspects of business activities & the necessary problems which the business is facing.

- **Data understanding:**

It involves study of data, shape, datatypes, number of rows and columns, type of columns and categories them into numerical and categorical data.

• **Data preparation:**
This involves Preprocessing of Data
• Access the data
• Ingest (or fetch) the data
• Cleanse the data
• Format the data
• Combine the data
• And finally Analyze the data

- **Variable information:**

Variable Name	**Variable Description**
Customer ID	Primary key of the record.
Churn	Information about Churn of the Customers.
Monthly Revenue	Revenue of each Customer
Monthly Minutes	Number of Minutes call spoken by Customer
Total Recurring Charge	The Charges for the Service
Director Assisted Calls	When we call an operator to request a telephone number
Overage Minutes	Count of Call used over duration to particular post-paid cell phone plan
Roaming Calls	The ability to get access to the Internet when away from home at the price of a local call or at a charge considerably less than the regular long-distance charges.
Three way Calls	A way of adding a third party to your conversation without the assistance of a telephone operator.
Dropped Calls	Count of Phone calls gets disconnected somehow from the cellular network.
Blocked Calls	Count of Telephone call that is unable to connect to an intended recipient.
Unanswered Calls	Count of Calling that an individual perceives but is not currently pursuing.

Variable Name	Variable Description
Received Calls	Number of calls received by the customer.
Out bound Calls	Call initiated by the call centre agent to customer on behalf of client to know the target customer behaviour and needs.
Inbound Calls	In inbound calls, call-centre or customer-care receives call from customer with issues and questions.
Peak Calls In Out	Amount of time period with fewer calls than are handled in a busy period.
Call Forwarding Calls	Count of Calls Forwarded by user.
Dropped Blocked Calls	Number of VM messages customer currently has on the server.
Call Waiting Calls	Duration of call-in waiting period
Months In Service	Number of months customer using service.
Unique Subs	subscription of different networks
Active Subs	subscriptions of the networks that are active or in usage.
Service Area	Network service area
Handset Models	Count of Handsets are used to Contact one to one.

Age HH1	User aged below 45
Age HH2	User aged above 45
Children in HH	Whether there are Children in House hold
Handset Refurbished	Are the handsets refurbished or not
Handset Web Capable	Are the handsets capable of internet connectivity
Truck Owner	Is the user a Truck Owner
RV Owner	Is the user an RV owner
Home Ownership	Is the house the user is staying, his own
Buys Visa Mail Order	Does the user buy Visa Mail order
Responds to Mail Offers	Does the user respond to Mail offers
Opt-out Mailings	Did he opt out of the mail offers sent to him

Non-US-Travel	Does the user travel to other countries
Owns-Computer	Does he have a computer or not
Has-Credit Card	Does he have a credit card or not
Retention Calls	No of Retention Calls
Retention Offers Accepted	Customers accepting retaining the retaining offers given by the company.
New Cell phone User	Number of customers buying new cell phone.
Referrals Made By Subscriber	Referrals made by the existing customer to the other customer.
Income Group	The column talks about the customer saying to which category the customer belongs to.
Owns Motorcycle	The columns ask about the customer weather the customer owns a motorcycle or not.
Adjustments To Credit Rating	Rating Scale
Handset Price	Its amount paid by the customer for his cell phone.

- **Modeling**

Based on the observation of Descriptive & Inferential Statistic & recognizing the right model.

- **Evaluation**

Uses some metric or combination of metrics to "measure" objective performance of model. Test the model against previously unseen data.

- **Deployment**

Applying the Data to the model

Reference:

- Customer Churn Analysis in Telecom Industry Dataset :- https://www.kaggle.com/datasets/jpacse/datasets-for-churn-telecom

- Customer Churn Analysis

Brief Overview of Customer Churn Analysis and Prediction with Decision Tree Classifier.Retrieved from https://towardsdatascience.com/customer-churn-analysis-4f77cc70b3bd

- Customer churn analysis: Churn determinants and mediation effects of partial defectionin the Korean mobile telecommunications service industry (2006). Retrieved from http://people.stern.nyu.edu/shan2/customerchurn.pdf

- Teemu Mutanen.Customer churn analysis – a case study.

Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.7169&rep=rep1&type=pdf

- Churn Analysis: 3-Step Guide to Analyzing Customer Churn Dominique Jackson(March 31, 2020).
Retrieved from https://baremetrics.com/blog/churn-analysis

- Customer Churn Analysis: A Comprehensive Guide Amit Phaujdar on ChurnAnalysis, Marketing Analytics (March 15th, 2021).
Retrieved from https://hevodata.com/learn/understanding-customer-churn-analysis/

- Understanding Random Forest

How the Algorithm Works and Why it Is So Effective.

Retrieved from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

- Logistic Regression.

Retrieved from https://www.sciencedirect.com/topics/computer-science/logistic-regression

- Decision Tree Algorithm, explained.

Retrieved from https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

- Prashant Gupta : Decision Trees in Machine Learning(May 18, 2017).

Retrieved from https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

- Logistic regression.

Retrieved from https://www.ibm.com/topics/logistic-regression

- Gradient Boosting from scratch.

Retrieved from https://blog.mlreview.com/gradient-boosting-from-scratch-1e317ae4587d

Notes For Project Team

Original owner of data	PAMINA
Data set information	The Data Contains information about Telecom Company
Any past relevant articles using the dataset	NA
Reference	Telecom Churn (Cell to Cell)
Link to web page	https://www.kaggle.com/datasets/jpacse/datasets-for-churn-telecom

****