**ONLINE RETAIL CUSTOMER SEGMENTATION**

# Gaurav Singh

**AlmaBetter,Bangalore**

## ABSTRACT:

In this cutting-edge international of enterprise, customer Churn is one of the foremost worries for various commercial enterprise owners or the corporations for preserving present and attracting new clients. evaluation of diverse sorts of clients may be carried out with the aid of gaining knowledge of purchaser relationship management which in turn provides sturdy assist for business decisions.

customer churn happens whilst certain customers are now not dependable or a part of a specific enterprise. losing clients will no longer handiest bring about losses but additionally broaden a risk to the company. because of multiple competition in the equal business, the re-engagement of customers who're less fascinated is critical instead of enticing a brand new one. it is found that obtaining new shoppers is dearer than keeping the existing customer. Churn prediction is a new promising method in customer relationship management to investigate consumer behavior through figuring out customers with a high chance to discontinue the employer based totally on studying their past statistics and also perceive techniques for development.

once a purchaser turns into a churn, the loss incurred with the aid of the corporate isn't always just the lost sales however also the charges involved in extra advertising on the way to appeal to new clients. decreasing consumer churn is a key commercial enterprise intention. This dataset carries statistics of transactions that took place among December 1, 2010 and December 1,2011. this is recorded

from a web retail present store based totally inside the united kingdom. here segmentation of clients has been carried out through the use of RFM method and k-approach set of rules. At ultimate, the client churn prediction is achieved using logistic regression, Random wooded area and XGBoost classifier to predict the churning conduct of the customers. The proposed approach is useful to assess customer loyalty and to manage customer relationships in an effective manner.

***Keywords: Customer segmentation, exploratory data analysis, Data Cleaning, clustering, Prediction, recency, frequency, monetary value, logistic regression, random forest, Xgboost.***

## INTRODUCTION:

## In latest years, with the fast development of electronic commerce, the numbers of e- commerce groups are booming increasingly and on a big scale, and the service has come to be more and more homogeneous, making opposition extra intense among e-commerce organizations. underneath an ecommerce environment, businesses use the internet platform to service clients, customers browse the community platform, the shopping for manner produces a huge quantity of information traffic, and the visitors within the form of information is straightforward to get admission to for any ecommerce agencies. based on the precise benefits of ecommerce groups, huge quantities of statistics thru records mining, statistics wishes to get clients, offer customers with personalized service, and continuously enhance consumer pride and loyalty, which has turn out to be the main intention of the e-trade international.

## Inside the retail region, a few customers stick around whilst others prevent purchasing at a particular keep after a positive time frame. Detecting which clients have decided to buy some place else and which of them are idle at the immediately, may be a tough undertaking to any agency.

## A client in all likelihood to break the relationship or lower the purchase price is called churn.

## There are many motives for client churn. depending on which purchaser churn can be divided into two classes: energetic and passive. active churn, namely voluntary leaving. manner a purchaser does not do online shopping because of his/her personal reasons which include changing jobs, pleasant of provider, commercial enterprise competition, loss of expert etc. Passive churn, additionally called involuntary leaving, refers to the type of customer churn that the employer ought to be answerable for. This occurs due to the fact the company makes a decision to cancel customers' money owed for a few motive, usually due to their credit score troubles.

## Client churn is the tendency of clients to prevent purchasing with a corporation over a term. patron churn is also referred to as client attrition or patron defection. Churning reduces increase. therefore, organizations should have a right described approach to compute consumer churn rate for a given time. through retaining tune of churn fee, organizations are regularly equipped to succeed in terms of client retention. retailers need an excellent method to manipulate consumer churn. Measuring the churn price is form of important for retail corporations due to the fact the metric reflects purchaser response in the direction of the provider, pleasant, charge and competition. Churn prediction envisions the likelihood of clients to churn. It pares the funding on gaining new clients and enables to retain the prevailing consumer. The marketing efforts and amount spent on attracting a new customer is better and more hard than clinging to current clients. customers who are unlikely to make a buy or willing to shift the purchasing web page due to cautiousness with money, looking forward to fashionable and assortment in products may be convinced and clutched. The customers who're ending the connection because of precious and unavoidable motives are free to leave.

## end result is company, even though we put money into involuntary churners. goal advertising aids to attain the customers and hook up with them. The voluntary

## Churners are frequently stopped by means of extending reductions, amending the goods to clients desire and via sending out trigger mails. Concentrating only on voluntary churners will scale back the fee of presenting benefits to yet and all churned customers.

As the development of the web retail marketplace intensifies opposition the various enterprise. For the e-trade industry, the customer churn price is excessive, business operators need to don't forget how to decrease the customer churn price of on-line shopping. because the patron's conduct is predictable, via the relevant statistics collected to hold out the relevant evaluation can discover the purchaser's destiny trading dispositions. For business operators, to reduce the quantity of misplaced clients, an effective manner is to find the client who has the free tendency and do the applicable pre-manage paintings. In current years, online buying consumer churn prediction has become a vital path of e-trade business research.

## Problem Statement:

In this project, your task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK- based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

**Data Set Information:**

**Invoice No:** Invoice number. Nominal,a 6-digit

integral number uniquely assigned to each

transaction. If this code starts with

letter 'c', it indicates a cancellation.

**StockCode:** Product (item) code. Nominal, a 5-

digit integral number uniquely assigned to

each distinct product.

**Description:** Product (item) name. Nominal.

**Quantity:** The quantities of each product

(item) per transaction. Numeric.

**Invoice Date:** Invoice Date and time.

Numeric, the day and time when each

transaction was generated.

**UnitPrice:** Unit price. Numeric, Product

price per unit in sterling.

**CustomerID:** Customer number. Nominal,

a 5- digit integral number uniquely

assigned to each customer.

**Country**: Country name. Nominal, the

name of the country where each customer

**Methodology:**

Although the data used must be kept private, it is critical to compare the results with other conventional machine learning algorithms to demonstrate the significance of the positives and negatives of each method considered in this study. This section explains the algorithms used in this study in detail. I’ve done preprocessing, handled null values, created required features, done Exploratory data analysis, done some feature engineering, created RFM model(recency, frequency and monetary value),created clusters, and at last done prediction using classification.

Furthermore, there is no single machine-learning algorithm that must be applied optimally in every scenario. As a result, three prediction

**STEPS EVOLVED IN DATA CYCLE**

## DATA ANALYSIS:

1. **DATA SOURCING**

## DATA PREPROCESSING:

1. **DATA CLEANING**

## DATA TRANSFORMATION:

1. **DATA DEDUPLICATION:**

## MISSING VALUES:

1. **DROPPING MISSING VALUES:**
2. **HANDLING OUTLIERS:**

# ALGORITHMS:

1. **LINEAR REGRESSION:**

Linear regression is a supervised machine learning model mostly used in forecasting. Supervised machine learning models are those where we use training data to build a model and then test the accuracy of the model using a loss function.

Linear regression is one of the most well- known time series forecasting techniques used for predictive modeling. As the name suggests, it assumes a linear relationship between a set of independent variables

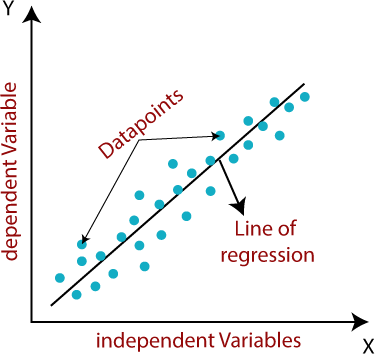
and a dependent variable (variable of interest).

EQUATION OF BEST FIT LINE.

y = β0 + β1x

to our data. Here x is called the independent variable or predictor variable and y is called the dependent variable or response variable. Before we talk about how to perform the fit, let's take a closer look at the important quantities from the fit:

* β1 is the slope of the line: this is one of the most important quantities in any linear regression analysis
* β0 is the intersection of the line.



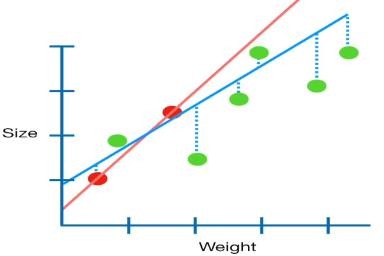
# RIDGE REGRESSION:

Ridge regression is a model tuning method used to analyze any data that suffers from multicollinearity. This method performs L2 regularization.

When the problem of multicollinearity occurs, the least squares are unbiased and the variances are large, resulting in the predicted values being far from the true values.

we came to the conclusion that we would like to reduce the complexity of the model, i.e. the number of predictors.

We could use forward or backward selection to do this, but that way we wouldn't be able to say anything about the effect of the removed variables on the response. Removing predictors from the



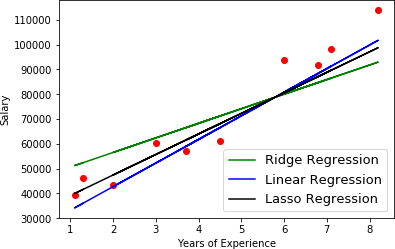
model can be thought of as setting their coefficients to zero. Instead of forcing them to be exactly zero, let's penalize them if they are too far from zero, thus forcing them to be small all the time.

This way we reduce the complexity of the model while keeping all the variables in the model. This is essentially what Ridge Regression does.

# LASSO REGRESSION:

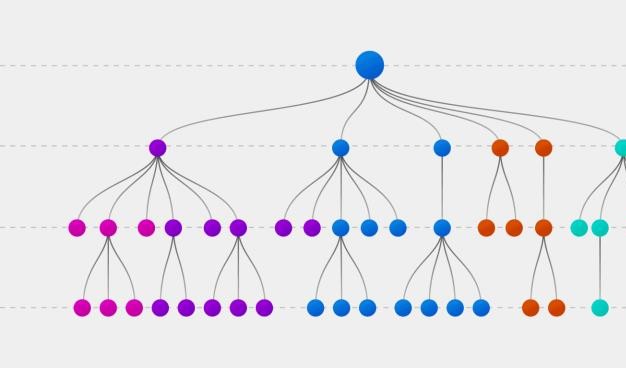
Lasso, or the least absolute shrinkage and selection operator, is conceptually quite similar to ridge regression. It also adds a penalty for nonzero coefficients, but unlike ridge regression, which penalizes the sum of the squares of the coefficients (the so- called L2 penalty), lasso penalizes the sum of their absolute values (the L1 penalty). As a result, for high values of λ, many coefficients are exactly zero under the lasso, which is never the case for ridge regression.

The only difference in the comb and lasso loss functions is in the penalty conditions. Under the lasso, loss is defined as:



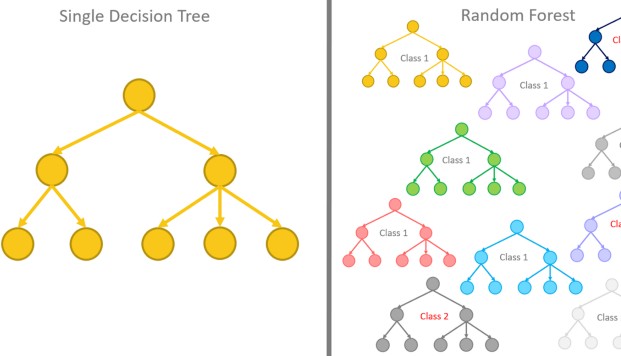
# DECISION TREE:

A decision tree is the most powerful and popular tool for classification and prediction. A decision tree is a tree structure similar to a flowchart, where each internal node denotes a test on an attribute, each branch represents the result of the test, and each leaf node (terminal node) has a class label. A tree can be "learned" by dividing the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. Decision trees classify instances by ordering them in a tree from the root to some leaf node that provides the instance's classification. An instance is classified by starting at the root node of the tree, testing the attribute specified by that node, and then moving down the branch of the tree corresponding to the value of the attribute, as shown in the figure above. This process is then repeated for the subtree rooted at the new node.



# RANDOM FOREST:

Random Forest is a wrapper type of decision tree algorithm that creates a series of decision trees from a randomly selected subset of the training set, collects labels from those subsets, and then averages the final prediction depending on how many times the label has been used. predicted of all.



# GRADIENT BOOSTING:

The term gradient gain consists of two sub- terms, gradient and gain. We already know that gradient boosting is a boosting technique. Let's see how the term "gradient" relates here.

Gradient boosting redefines boosting as a numerical optimization problem where the objective is to minimize the loss function of the model by adding weak pupils using

gradient descent. Gradient descent is a first-order iterative optimization algorithm for finding the local minimum of a differentiable function. Since gradient boosting is based on the minimization of a loss function, different types of loss functions can be used, resulting in a flexible technique that can be applied to regression, multi-class classification, etc.

# CONCLUSIONS:

a.)So finally, we can conclude here!

Starting with loading the data so far, we have done Data cleaning, handled Null values, EDA, encoding of categorical columns, feature

b.) selection, model building and then

Evaluation of the model with

different evaluation metrics.

We've got the model accuracy in

the range of 66% to 90%.

c.) So, the accuracy of our best model

is 90% which is by XGBoost

so, Xgboost is our optimal model.

**REFERENCES:**

* + - <https://book.akij.net/eBooks/20> [18/May/5aef50939a868/Data\_S](https://book.akij.net/eBooks/2018/May/5aef50939a868/Data_Science_for_Bus.pdf) [cience\_for\_Bus.pdf](https://book.akij.net/eBooks/2018/May/5aef50939a868/Data_Science_for_Bus.pdf)
    - [https://bunker2.zlibcdn.com/dtok](https://bunker2.zlibcdn.com/dtoken/01c5fc197a94283bfb0c0943bd5b2d0c) [en/01c5fc197a94283bfb0c0943bd](https://bunker2.zlibcdn.com/dtoken/01c5fc197a94283bfb0c0943bd5b2d0c) [5b2d0c](https://bunker2.zlibcdn.com/dtoken/01c5fc197a94283bfb0c0943bd5b2d0c)
    - W3SCHOOLS
    - GEEKS FOR GEEKS
    - ALMA BETTER
    - PANDAS

Towards DataScience

Data Science from scratch

Neptune