**Online Retail Customer Segmentation**

**SUNANDA DEBNATH, Ajay Tiwari,**

**Data science trainees,**

**Alma Better, Bangalore**

**Abstract: -**

These days, you may customize everything. There’s no one-size-fits-all approach. But, for business, that is in reality a fantastic thing. It creates numerous areas for wholesome competition and possibilities for businesses to get innovative approximately how they gather and hold customers.

One of the essential steps towards higher personalization is customer segmentation. This is where personalization starts, and the right segmentation will assist you're making choices concerning new features, new products, pricing, advertising, and marketing strategies allows us to better understand our customers helping us target these customers in a more efficient manner and improve the customer experience.

.

## 1. Problem Statement

This project aims to identify major customers Customer segmentation is the process of segments on a transnational (extending or dividing your customers into sub-groups going beyond national boundaries based on shared features. transnational corporation.) data set which contains all the transactions occurring Because you use on-site data to optimize between 01/12/2010 and 09/12/2011 for advertising off-site, segmentation happens UK-based and registered non-store online after the fact, unlike customization and retail. The company mainly sells unique all-targeting.

Occasion gifts. Many customers of the because you need to build triggers so that company are wholesalers. Your consumers see the advertisements Ecommerce customer segmentation divides when they arrive, you need to do your clients into smaller groups who share targeting and personalization before they have a common interest, making it easier to up with offers and calls to action.

**Data Description:**(**Attribute Information)**

1. **Invoice No:**  Invoice number. Nominal, is a 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
2. **Stock Code:** Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct
3. **Description:** Product (item) name. Nominal.
4. **Quantity:** The quantities of each product (item) per transaction. Numeric.
5. **Invoice Date:** Invoice Date and time. Numeric, the day and time when each transaction was generated.
6. **Unit Price:** Unit price. Numeric, Product price per unit in sterling.
7. **Customer ID:** Customer number. Nominal, is a 5-digit integral number uniquely assigned to each customer.
8. **Country:** Country name. Nominal, is the name of the country where each customer resides.

## 2. Introduction

When a customer makes a purchase, there is some information that is now stored. Such information is given below:

In e-commerce, customer segmentation InvoiceNo, StockCode, Description, refers to the use of customer data to divide Quantity, InvoiceDate, UnitPrice, customers into groups that share the same CustomerID, Country. behavior and characteristics such as gender, taste or shopping patterns, interests, and more. Segmenting the customer base helps in better understanding the customers and thus personalizing marketing and communication for each segment. This is very beneficial because people tend to respond better and be

of greater value to your business when they

feel their needs and interests are being specifically addressed.

## 3. Steps involved:

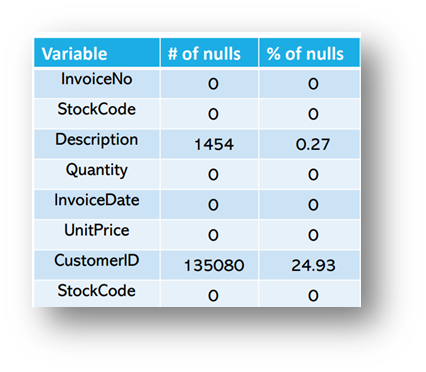
* **Exploratory Data Analysis**

(EDA) is utilized by statistics scientists to investigate and inspect statistics units and summarize their major characteristics, frequently using statistics visualization methods. It enables deciding how exceptional to govern statistics reasserts to get the solutions you need, making it less complicated for statistics scientists to find out patterns, spot anomalies, take a look at a hypothesis, or take a look at an assumption

After loading the dataset we performed. This process helped us figure out various aspects and relationships among independent/ feature variables. It gave us a better idea

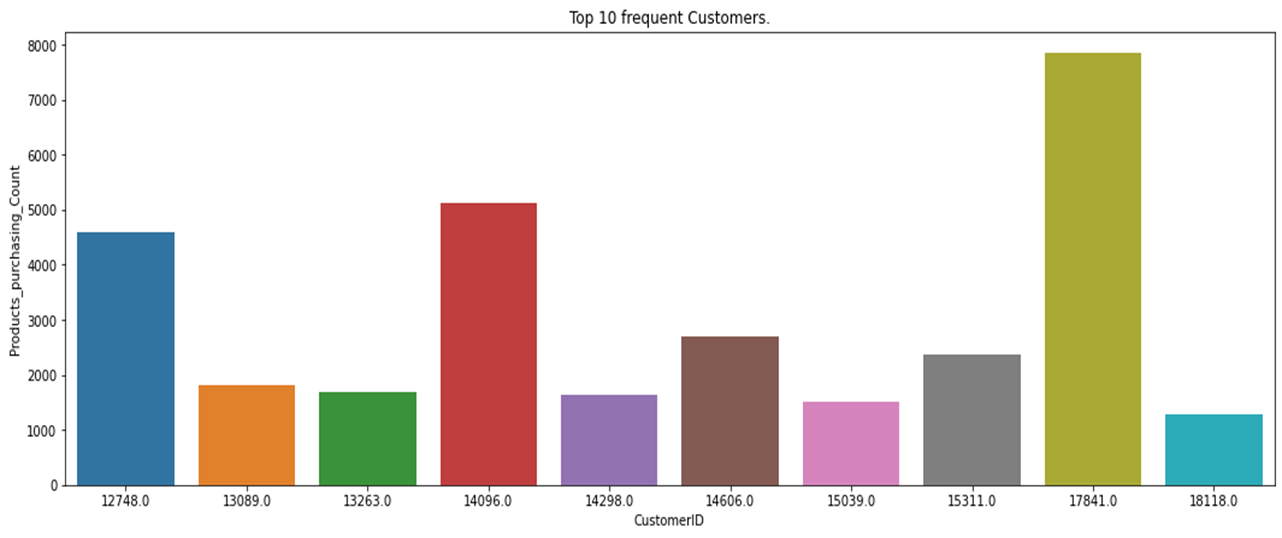
● **Null values Treatment**

Our dataset contains some null values like customer\_id, AND Description which may tend to our project so will come good at the beginning of our project in order to get a better resul**t**.

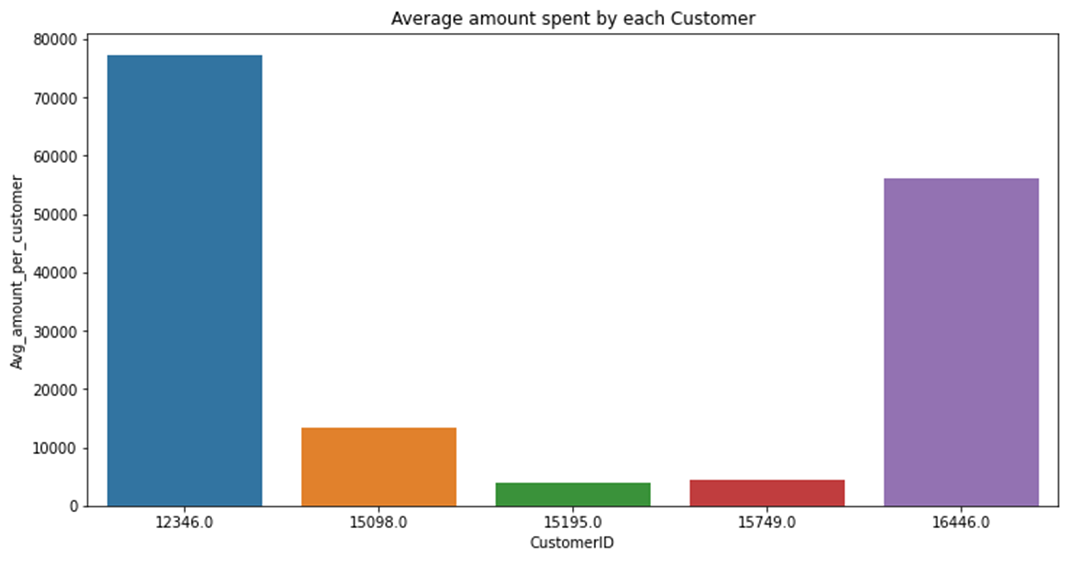


**m**

●  **Data pre-processing and transformation**

****

in this graph we can observe that based on purchasing count which are the top 10 most frequent customers.



The above graph represents

● **Standardization of features**

**Create the RFM model (Recency, Frequency ,Monetary value) for clustering made easy**

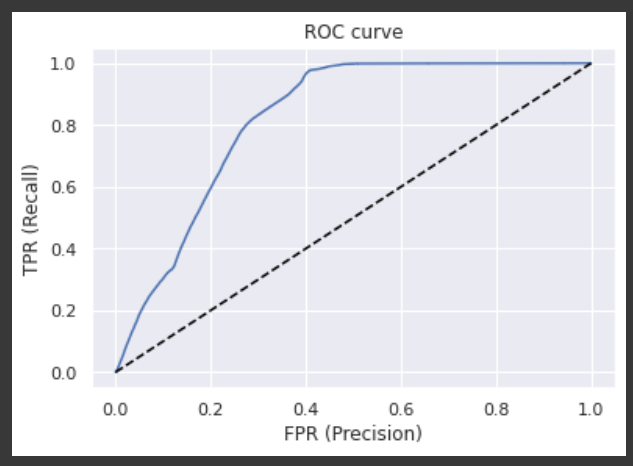
**Fitting different models**

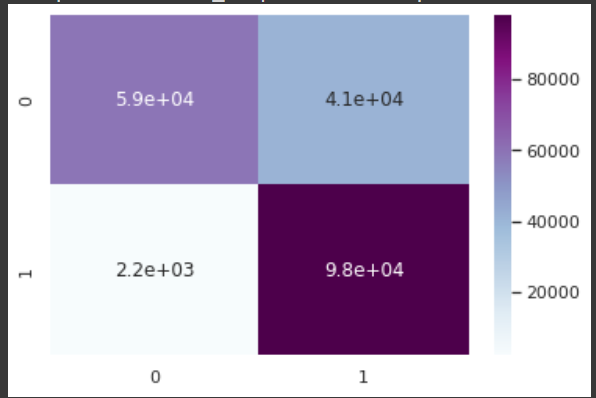
For modelling we tried various clustering algorithms like:

* **Model building, Predictions, and Forecasting**

### Logistic Regression:

Logistic regression is a sort of easy system learning that plays regression estimation on proportional, proportional, or specific data. Compared to more superior type and regression techniques, it's miles primarily based totally on a totally easy theory, however, probabilistic evaluation is viable for specific data. In different words, it's miles viable to expect the prevalence of a particular occasion via way of means of the use of an unbiased variable that has an immediate effect on the established variable. Logistic regression evaluation can provide an explanation for the forms of institutions and interactions resulting from the version shape and might evaluate the effect of explanatory variables on reaction values via parameter inference. In addition, due to the fact it's miles viable to carry out discrimination and type primarily based totally on anticipated chance, numerous industries along with medicine, telecommunications, and finance are acting duties to expect the chance of an occasion going on the use of logistic regression evaluation.





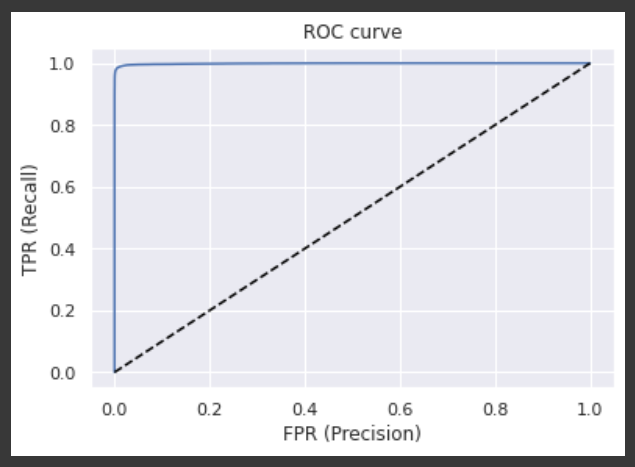
### 

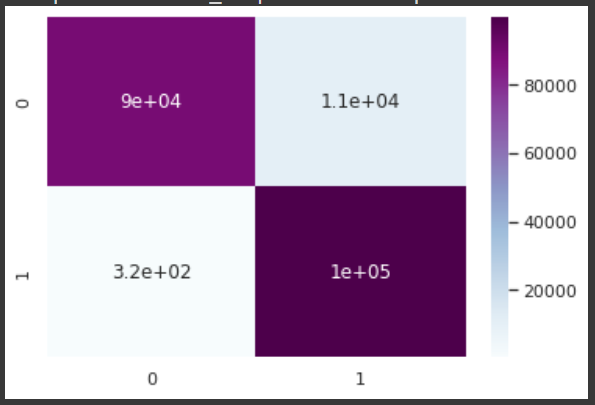
### Random Forest :

Random Forest is a famous machine learning algorithm that belongs to the supervised learning technique. It may be used for each Classification and Regression issue in ML. It is primarily based totally on the idea of ensemble getting to know, that's a procedure of mixing a couple of classifiers to remedy complicated trouble and enhance the overall performance of the model.

As the call suggests, "Random Forest is a classifier that carries some of the choice bushes on diverse subsets of the given dataset and takes the common to enhance the predictive accuracy of that dataset." Instead of counting on one choice tree, the random wooded area takes the prediction from every tree and is primarily based totally on the bulk votes of predictions, it predicts the very last output.

The extra quantity of bushes withinside the wooded area results in better accuracy and forestalls the trouble of overfitting.



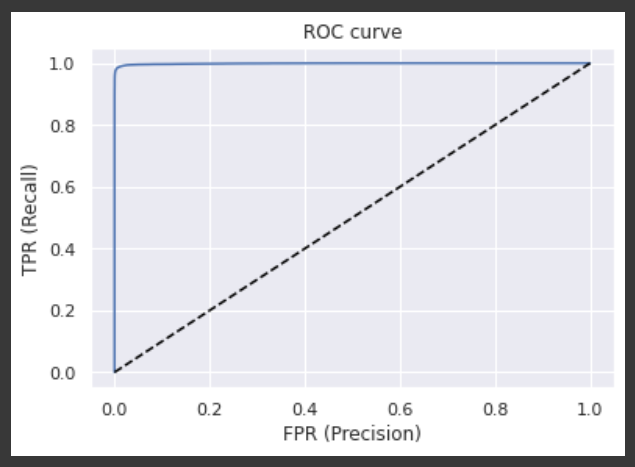


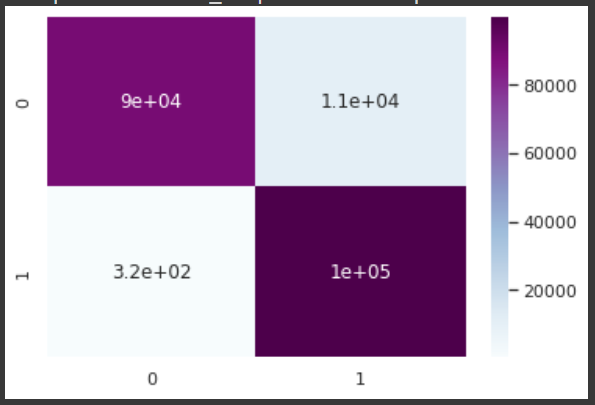
### XGB classifier :

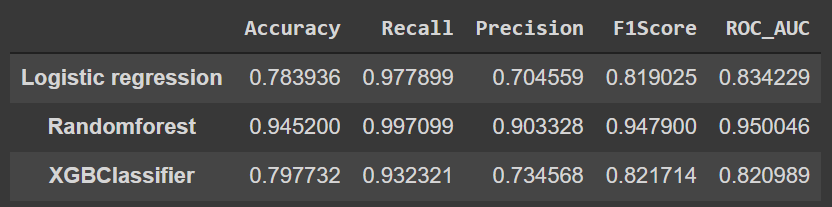
XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides a parallel tree boosting and is the leading machine-learning library for regression, classification, and ranking problems.

It’s vital to an understanding of XGBoost to first grasp the machine learning concepts and algorithms that XGBoost builds upon: supervised machine learning, decision trees, ensemble learning, and gradient boosting.

Supervised machine learning uses algorithms to train a model to find patterns in a dataset with labels and features and then uses the trained model to predict the labels on a new dataset’s features.





**Combining all regression models: **

## 6. Challenges:

## Identify a highly imbalanced data set and manage it carefully.

## Ensure that the model treats all groups fairly. Set prediction thresholds before deploying the model. After deployment, understand the behavior of the model on real data. It is difficult to perform predictive analysis for end users.

## 7. Conclusion:

87.7% of clients spoke back as No to purchasing car insurance. It surely indicates that a maximum of the clients aren't interested in shopping for car insurance.

• Males are 30% more likely to reply as sure for car coverage than females. So business enterprises may want to attention greater on concentrated on male clients and do greater promotions centered toward the woman clients.

• Most of the clients have riding licenses and out of them, 12% are probable to reply as sure for car coverage.

• There isn't any factor in attaining out to clients who have already got car coverage as nearly they all spoke back negatively about purchasing every other coverage.

• 22% of clients who spoke back definitely do not have preceding coverage. So, the business enterprise ought to attention greater such clients as conversion opportunities are better in such cases.

• Company ought to attention to clients whose car is greater than 2 years old, as 30% of instances they may be interested in shopping for coverage, which is large in comparison to different features.

• Customers with cars aged much less than 12 months are least interested in insurance as at the same time as shopping for a car human beings frequently purchase 1 year coverage. Companies should not spend greater time on those clients as simply 4% of instances they may be probable to mention Yes for car insurance.

• Customers who have broken their motors in beyond are greater sensitive toward shopping for car insurance. In truth 24% of instances, they spoke back definitely primarily based totally on this dataset.

**7. References**

1. Stack Overflow!
2. GeeksforGeeks
3. Analytics Vidhya
4. Almabetter
5. GitHub
6. Towards data science

# **Conclusion:-**

* Due to the Response variable's value 1 being much lower than its value 0, the provided dataset is an imbalanced dataset.
* Compared to their female counterparts, male consumers own a little bit more vehicles and have a higher likelihood to get insurance.
* Customers between the ages of 30 and 60 are the most likely to get insurance whereas Vehicle insurance is not interesting to anyone under the age of 30. The lack of involvement, a lack of knowledge about insurance, and possibly the lack of expensive vehicles are potential causes.
* Customers with driving licenses are more likely to purchase insurance
* Compared to consumers with vehicles less than one-year-old, those with vehicles between one and two years old are more interested in purchasing insurance.
* Due to their personal experience with the costs associated with vehicle repairs, customers with vehicle damage are more likely to purchase insurance.
* The variable such as Age, previously insured, and Annual premium is more affect the target variable.

# We used different types of algorithms to train our model like Logistic Regression, Random Forest model, Decision tree, and XGB Classifier. And Also, we tuned the parameters of the XGB Classifier and Random Forest model. Comparing the model on the basis of precision, recall, accuracy, and F1 score we can see that the XGB Classifier model performs better. Even comparing the ROC curve XGB Classifier performed better because curves closer to the top-left corner indicate