CUSTOMER SEGMENTATION

SUVENDU NAYAK

Data science trainee,

Alma Better, Bangalore.

Abstract:

In this current world of business, Customer Churn is one of the major concerns for various business owners or the organizations for maintaining existing and attracting new customers. Analysis of various types of customers can be conducted by researching customer relationship management which in turn provides strong support for business decisions. Customer churn occurs when certain customers are no longer loyal or a part of a particular business. Losing customers will not only result in losses but also develop a threat to the organization. Because of multiple competitors in the same business, the re-engagement of customers who are less interested is essential rather than engaging a new one. It is observed that acquiring new buyers is costlier than retaining the present customer. Churn prediction is a new promising method in customer relationship management to analyse customer behaviour by identifying customers with a high probability to discontinue the company based on analysing their past data and also identify strategies for improvement. Once a customer becomes a churn, the loss incurred by the corporate isn't just the lost revenue but also the prices involved in additional marketing in order to attract new customers. Reducing customer churn is a key business goal. This dataset contains records of transactions that happened between December 1, 2010 and December 1, 2011. This is recorded from a web retail gift store based in the United Kingdom. Here segmentation of customers has been done by using RFM technique and K-means algorithm. At last, the customer churn prediction is using logistic regression, Random forest and XGBoost classifier to predict the churning behaviour of the customers. The proposed approach is helpful to assess customer loyalty and to manage customer relationships in an effective manner.

Keywords: Customer segmentation, clustering, Prediction, recency, frequency, monetary value, logistic regression, random forest, Xgboost.

1.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 UKbased and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

1.1. 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.

Stock Code: Product (item) code. Nominal, a 5digit 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 5digit integral number uniquely assigned to each customer.

Country: Country name. Nominal, the name of the country where each customer resides.

2. Introduction:

In recent years, with the rapid development of electronic commerce, the numbers of ecommerce businesses are booming more and more and on a large scale, and the service has become increasingly homogeneous, making competition more intense among e-commerce businesses. Under an ecommerce environment, companies use the Internet platform to service customers, customers browse the network platform, the buying process produces a large amount of data traffic, and the traffic in the form of data is easy to access for any ecommerce businesses. Based on the unique advantages of ecommerce businesses, large amounts of data through data mining, information needs to get customers, provide customers with personalized service, and constantly improve customer satisfaction and loyalty, which has become the main goal of the e-commerce world.

In the retail sector, some customers stick around while others stop shopping at a particular store after a certain period of time. Detecting which customers have decided to buy elsewhere and which of them are idle at the instant, may be a difficult task to any organization. A customer likely to break the relationship or lower the purchase rate is known as churn. There are many reasons for customer churn. Depending on which customer churn can be divided into two categories: active and passive. Active churn, namely voluntary leaving. Means a customer does not do online shopping due to his/her own reasons such as changing jobs, quality of service, business competition, loss of professional etc. Passive churn, also known as involuntary leaving, refers to the type of customer churn that the enterprise should be responsible for. This occurs because the enterprise decides to cancel customers' accounts for some reason, usually due to their credit problems.

Customer churn is the tendency of customers to stop purchasing with a company over a time period. Customer churn is also called customer attrition or customer defection. Churning reduces growth. Therefore, companies should have a proper defined method to compute customer churn rate for a given time. By keeping track of churn rate, organizations are often equipped to succeed in terms of customer retention. Retailers need a good strategy to manage customer churn. Measuring the churn rate is kind of crucial for retail businesses because the metric reflects customer response towards the service, quality, price and competition. Churn prediction envisions the likelihood of customers to churn. It pares the investment on gaining new customers and helps to retain the existing customer. The marketing efforts and amount spent on attracting a new customer is higher

and more difficult than clinging to existing customers. Customers who are unlikely to make a purchase or willing to shift the shopping site because of cautiousness with money, expecting standard and assortment in products can be convinced and clutched. The customers who are ending the relationship due to valuable and unavoidable reasons are free to leave.
Result is firm, though we invest in involuntary churners. Target marketing aids to reach the customers and connect with them. The voluntary churners are often stopped by extending discounts, amending the products to customers choice and by sending out trigger mails. Concentrating only on voluntary churners will scale down the cost of offering benefits to yet and all churned customers.

As the development of the online retail market intensifies competition among the industry. For the e-commerce industry, the customer churn rate is high, business operators need to consider how to minimize the customer churn rate of online shopping. Because the customer's behavior is predictable, through the relevant data collected to hold out the relevant analysis can find the customer's future trading tendencies. For business operators, to reduce the number of lost customers, an effective way is to find the customer who has the loose tendency and do the relevant pre-control work. In recent years, online shopping customer churn prediction has become a crucial direction of e-commerce business research.

3. 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 models. Furthermore, there is no single machine-learning algorithm that must be applied optimally in every scenario. As a result, three prediction algorithms were considered in this study to compare their performance.

1. Data Preprocessing.
2. Exploratory data analysis.
3. Feature engineering.
4. RFM model.
5. KMeans Clustering.
6. Build classification models.

1) Data preprocessing :

The dataset mentioned in the previous section is in the form of a CSV file. The CSV file is imported, cleaned, aggregated as per the requirements, and made into a new data frame. The retail sales dataset has a total size of 392692 tuples when compared to the original raw dataset which had around 541909 tuples. The preprocessing step involved cleaning of data by removing the NA values and also defining one more attribute called Total_price which is the product of Quantity and Unit price. This process also involved creation of two more datasets using aggregation which are customer aggregate data and invoice aggregate data. The missing entries in the customerID field, duplicate entries, cancelled transactions that completely got cancelled and the invoices which do not relate to customers are removed.

2) Exploratory data analysis :

Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modelling task. It is not easy to look at a column of numbers or a whole spreadsheet and determine important characteristics of the data. It may be tedious, boring, and/or overwhelming to derive insights by looking at plain numbers. Exploratory data analysis techniques have been devised as an aid in this situation.

Exploratory data analysis is generally cross classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate).

3) Feature engineering :

Feature engineering refers to a process of selecting and transforming variables when creating a predictive model using machine learning or statistical modeling (such as deep learning, decision trees, or regression). The process involves a combination of data analysis, applying rules of thumb, and judgement.

4) RFM model :

i) RFM technique :

RFM (Recency, Frequency, Monetary) analysis is a marketing model for customer segmentation. It is based on customer behaviour. It groups customers based on their transactional history that is how recently, how often and how much did they buy. RFM helps divide customers into various clusters to identify customers who are more likely to discontinue the business relationship.

1) Recency: The freshness of customer activities: Time since last transaction.

2) Frequency: The Frequency of customer transactions. E.g., the total number of recorded transactions.

3) Monetary: The total amount paid. E.g., the total transaction value.

RFM factors illustrate these facts:

a. The newer purchase, the more responsive the customer is to promotions.
b. The more frequently the customer buys, the more engaged and satisfied they are.
c. Monetary value differentiates heavy spenders from low value purchasers.

ii) RFM implementation :

Following are the steps for RFM Calculation –

Recency calculation- Using the most recent date in the complete dataset, the recency for all customers is calculated by subtracting the most recent date from the customers recent date of transaction. The most recent date from the complete dataset is found to 2011-12-09. This is used as a benchmark date for the calculation of recency value for each customer.

Frequency calculation- Adding the number of times a particular customer has purchased, gave the frequency.

Monetary calculation- Using the quantity bought in order and unit price of product, the total money spent by customer on each transaction is calculated as total. Sum of this total of each transaction is taken for each customer to get the total money spent by the customer.

CustomerID	Recency	Frequency	MonetaryValue	Recency_Q	Frequency_Q	MonetaryValue_Q	RFM_Segment	RFM_Score
12821	215	6	92.72	1	1	1	1.01.01.0	3
12831	263	9	215.05	1	1	1	1.01.01.0	3
12837	174	12	134.10	1	1	1	1.01.01.0	3
12864	139	3	147.12	1	1	1	1.01.01.0	3
12897	205	4	216.50	1	1	1	1.01.01.0	3

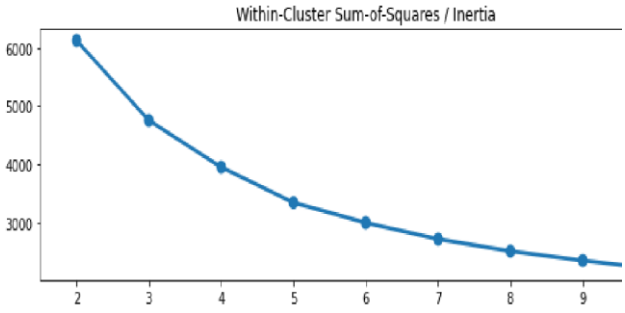
5) KMeans Clustering :

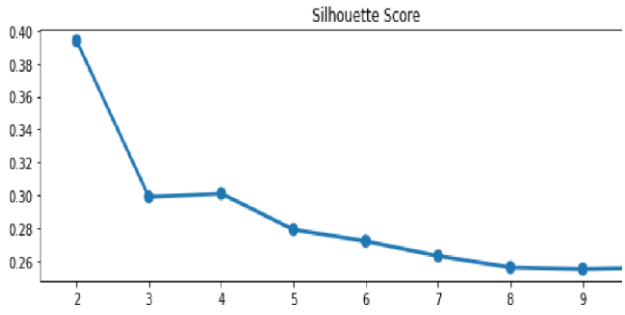
Clustering is defined as the process which divides the whole data into groups or clusters supporting the patterns within the data. To process the training data, the K-means algorithm in data processing starts with a primary group of randomly selected centroids, which are used as the beginning points for each cluster, then performs iterative calculations to optimize the positions of the centroids.

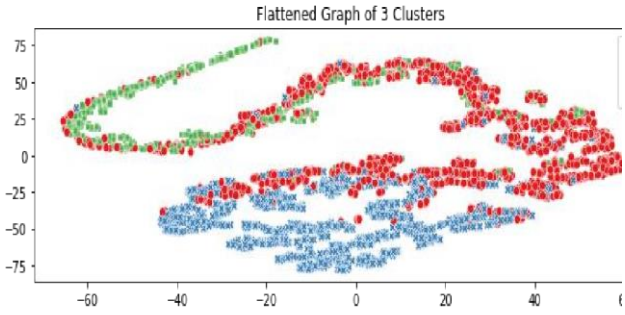
- Implementation of KMeans clustering :

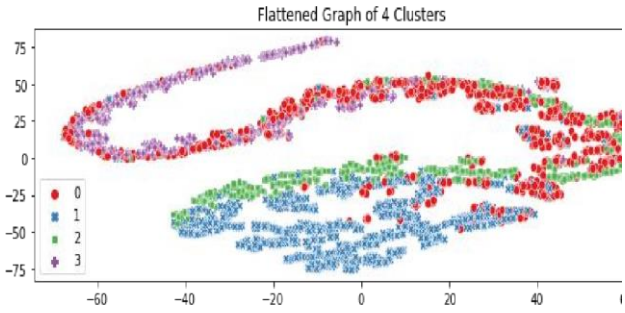
In this project, for creation of customer segmentation using K-Means algorithm based on the R, F, and M Scores, it is essential to decide the number of clusters to form i.e. the value of K. For deciding the value of k Elbow Technique is used.

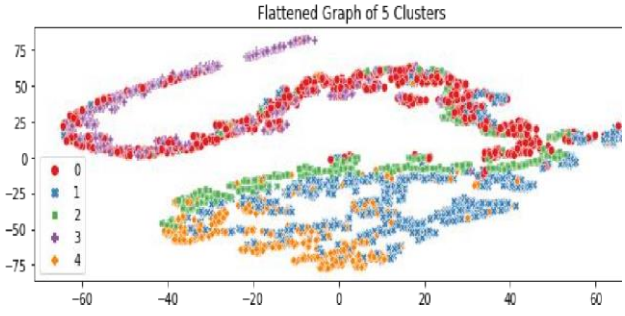
Elbow technique: The elbow method runs kmeans clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center and picking the elbow of the curve as the number of clusters to use. We can observe from Figure (5), as the number of clusters increases the sum of square distances are becoming lesser. And will take the count of clusters where this elbow is bending. In our case, the sum of square distance is dramatically decreasing at K = 3, so this is the optimal value to choose for no of clusters.











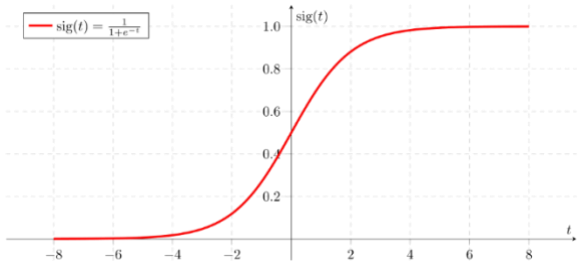
Cons:

6) Classification Models :

A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data. Feature: A feature is an individual measurable property of a phenomenon being observed. I've used 3 classification models i.e. Logistic regression, Random Forest and XGBoost and Random Forest is our Optimal model for further use.

4.1 Assumptions Of Models:

a) Logistic Regression:



Assumptions:

1. It assumes that there is minimal or no multicollinearity among the independent variables.
2. It usually requires a large sample size to predict properly.
3. It assumes the observations to be independent of each other.

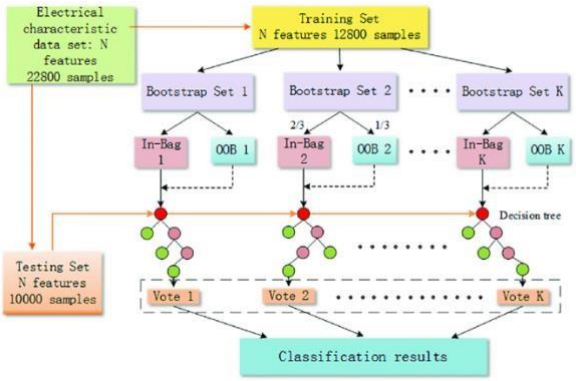
Pros:

1. Easy to interpret, implement and train. Doesn't require too much computational power.
2. Makes no assumption of the class distribution.
3. Fast in classifying unknown records.
4. Can easily accommodate new data points.
5. Is very efficient when features are linearly separable.

1. Tries to predict precise probabilistic outcomes, which leads to overfitting in high dimensions.
2. Since it has a linear decision surface, it can't solve nonlinear problems.

3. Tough to obtain complex relations other than linear relations.
4. Requires very little or no multicollinearity.
5. Needs a large dataset and sufficient training examples for all the categories to make correct predictions.

b) Random Forest:



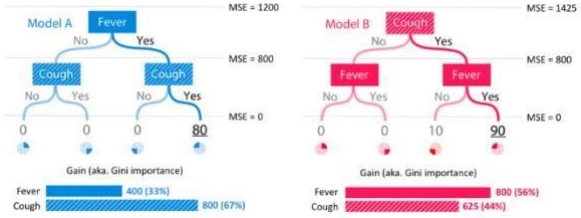
Assumptions:

1. Assumption of no formal distributions. Being a non-parametric model, it can handle skewed and multi-modal data. ***Pros:***

1. Robust to outliers.
2. Works well for non-linear data.
3. Low risk of overfitting.
4. Runs efficiently on large datasets. ***Cons:***

1. Slow training.
2. Biased when dealing with categorical variables.

c) XGBoost:



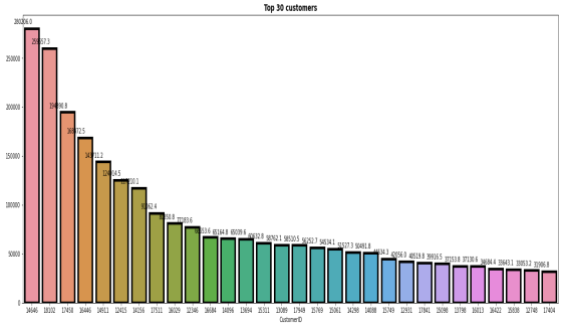
Assumptions:

1. It may have an assumption that the encoded integer value for each variable has an ordinal relation. ***Pros:***

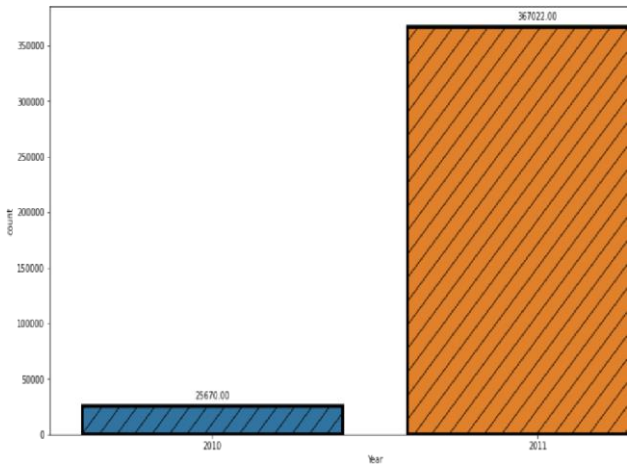
1. Can work in parallel.
2. Can handle missing values.
3. No need for scaling or normalizing data.
4. Fast to interpret.
5. Great execution speed. ***Cons:***

1. Can easily overfit if parameters are not tuned properly.
2. Hard to tune.

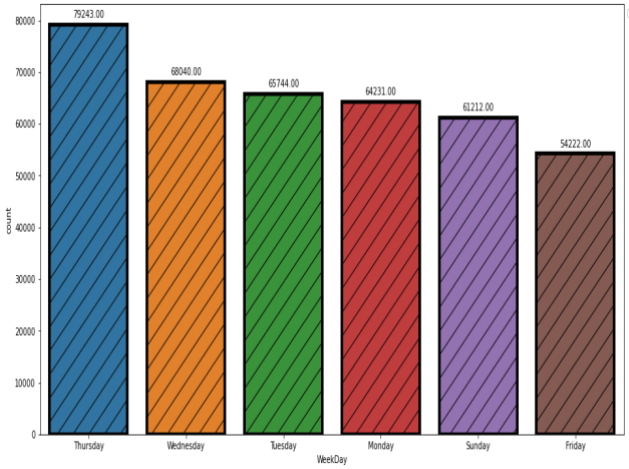
5. Trends:



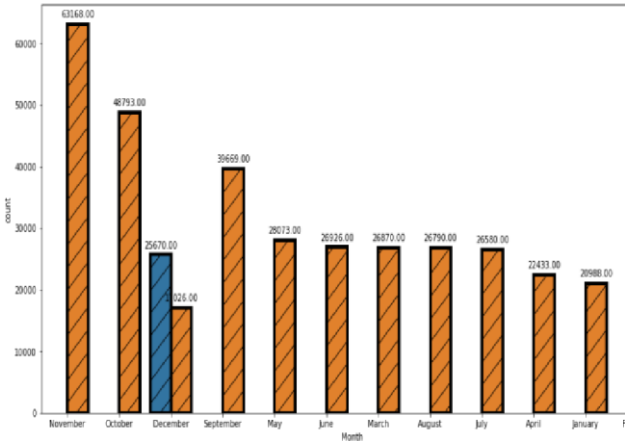
These are the top customers who buy often and spend good money.



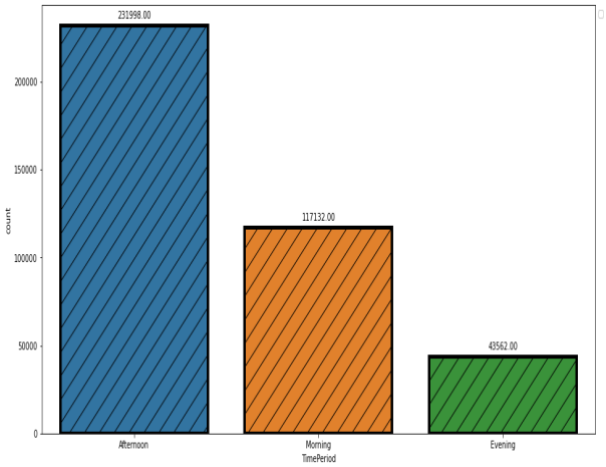
Here we can see a huge spike in 2011 this because we only have december months data from 2010.



Here we can see on Wednesday and Thursday there is more sale is



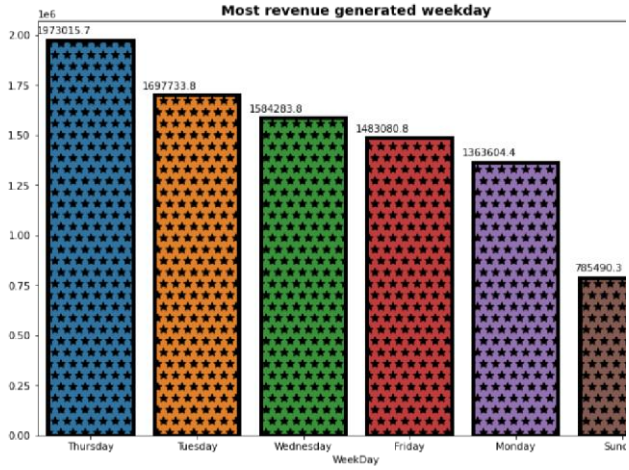
Here we can see there are huge sale in the month of October and November.



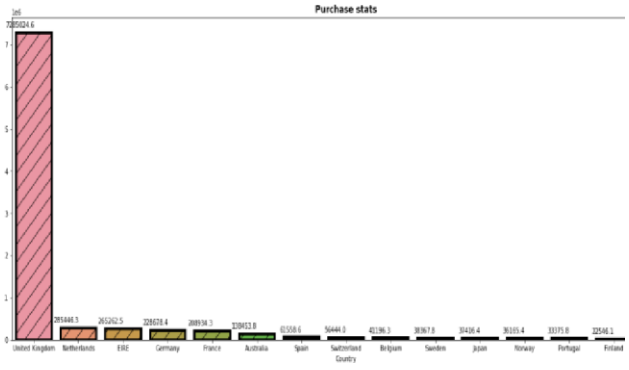
Here we can see people are buying more in the afternoon time period.

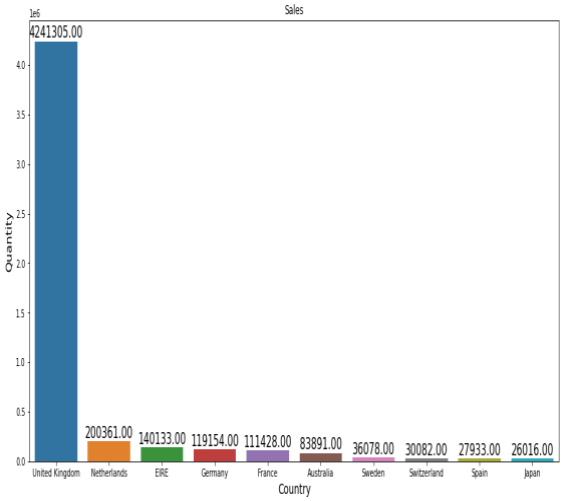
- Sales are very high in October, November & December
- We have only December sales data from 2010
- The Retail Store is Closed on Saturday as per the Information available
- More number of sales are happening during Middle of the week(Tuesday, Wednesday and Thursday in ascending order respectively)

- People are usually visiting the store during the afternoon hours when compared to Morning & Evening Times
- The total no of sales in December 2010 is higher than December 2011
- Management has to concentrate on the decrease of sales in December



On Thursday Company is generating the highest Revenue
On Sunday company is generating less revenue

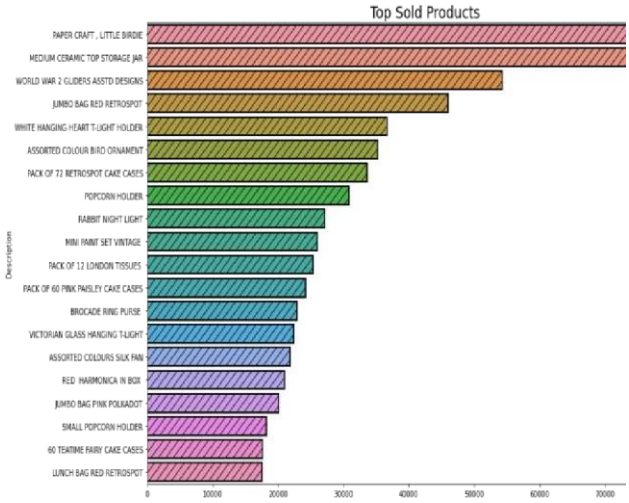




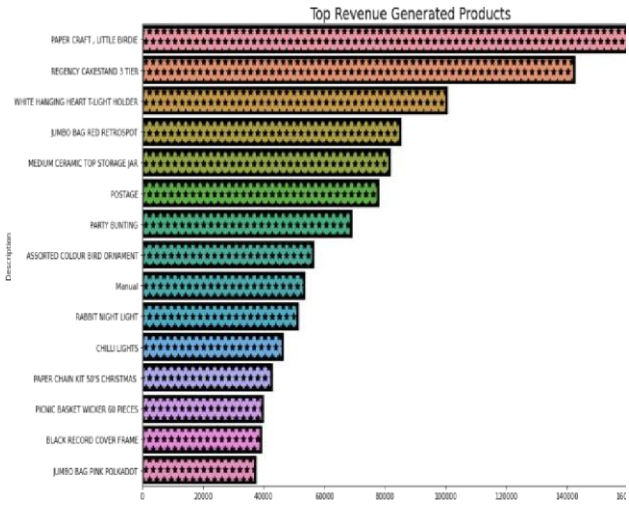
Above we can UK is the country which purchases in high quantity.

Above we can see the product sale categories.

- Here we can see the United Kingdom is the country which purchases more items as compared to other countries.
- Singapore is the least purchasing country.

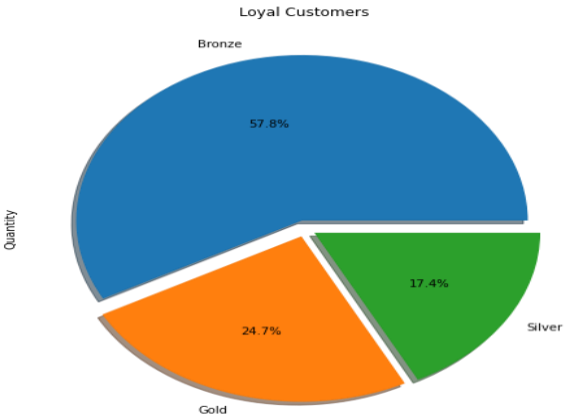


Here we can see the most sold product is Paper craft, little birdie and least sold product is lunch bad red retro spot.

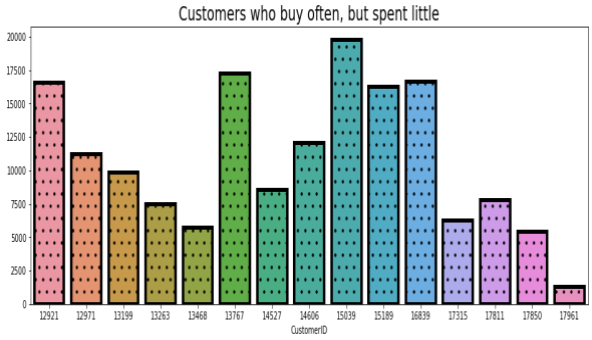


Here we can see the most revenue generated product is Paper craft, little birdie and least revenue generated product is jumbo bag pink polka dot.

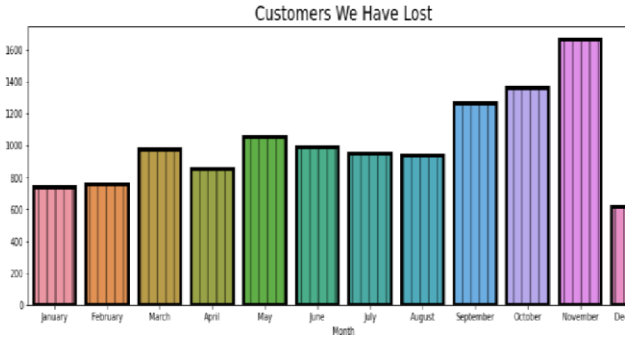
Above are the customer id who purchased the highest quantity of products.



Here we can see we have more no. of bronze customers.



Above we can see Customers who buy often, but spend little.



We can observe that a large number of customers are lost visiting the store. At the end of the year many customers are lost.

6. Steps involved:

- **Null values Treatment**

I've checked the null values in our dataset with isnull() function in heatmap plot there was many null values present in our dataset. Null values have been removed as the null values were present in customer id column and we have to do customer segmentation therefore customer id is necessary.

- **Exploratory Data Analysis**

I used this strategy after importing the dataset by comparing our different variables. This method assisted us in determining numerous characteristics and correlations between variables. It helped us understand which features behave in which ways in relation.

- **Feature Engineering**

I've created and selected important features which will help us in clustering and in creation of the RFM model.

- **Standardization of features**

Our main goal in this step was to scale our data into a standard format so that we could better use it for fitting and applying multiple algorithms. The main purpose was to ensure that specific behavior or processes within the chosen environment were consistent or uniform.

I've used minmax Scaler for scaling purpose.

- **RFM Model :**

The recency, frequency, monetary value (RFM) model is based on three quantitative factors namely recency, frequency, and monetary value. Each customer is ranked in each of these categories, generally on a scale of 1 to 5 (the higher the number, the better the result).

- **KMeans Clustering :**

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. Data points are clustered based on feature similarity. With the elbow method I came to know that the optimal cluster is 3.

- **Fitting Classification models**

For modelling we tried various classification algorithms like:

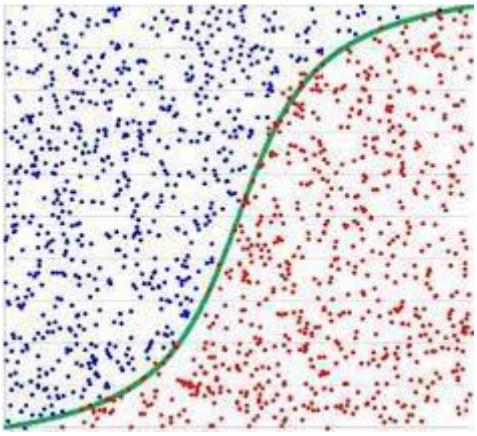
1. **Logistic Regression**
2. **Random Forest**
3. **XGBoost**

- **Tuning the hyperparameters for better accuracy**

  Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of tree-based models. like random forest and XGBoost.

7.1. Algorithms:

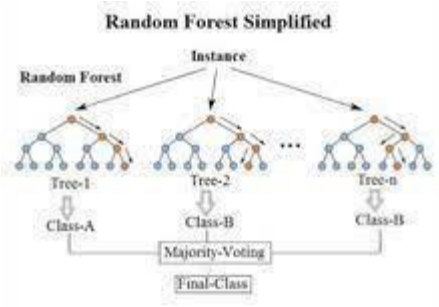
a) Logistic Regression:



Logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite like multiple linear regression, with the exception that the response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

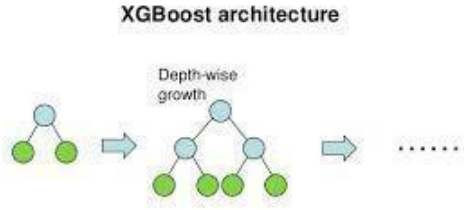
Mathematically, a logistic regression model predicts P(Y=1) as a function of X.

b) Random Forest:



A random forest is a supervised machine learning algorithm that is constructed from decision tree algorithms. This algorithm is applied in various industries such as banking and e-commerce to predict behavior and outcomes.

c) XGBoost:



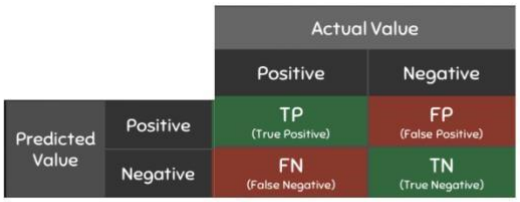
XGBoost is an algorithm that has recently been dominating applied machine learning for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance

7.2. Evaluation Metrics:

It is critical to acquire accuracy on training data, but it is equally critical to obtain a real and approximate answer on unknown data; otherwise, the model is useless. So, in order to construct and deploy a generalized model, we must evaluate it on many metrics, which allows us to better optimize, fine-tune, and acquire a better outcome.

There is no need for several measures if one is perfect. Because each evaluation measure fits on a distinct set of a dataset, it's important to grasp the pros and downsides of each.

a) Confusion Matrix:



- True Positive (TP) : Observation is positive, and is predicted to be positive.
- False Negative (FN) : Observation is positive, but is predicted negative.
- True Negative (TN) : Observation is negative, and is predicted to be negative.
- False Positive (FP) : Observation is negative, but is predicted positive.

Evaluation of the performance of a classification model is based on the counts of test records correctly and incorrectly predicted by the model. The confusion matrix provides a more insightful picture which is not only the performance of a predictive model, but also which classes are being predicted correctly and incorrectly, and what type of errors are being made. To illustrate, we can see how the 4 classification metrics are calculated (TP, FP, FN, TN), and our predicted value compared to the actual value in a confusion matrix is clearly presented in the below confusion matrix table.

b) Precision:

Is the ratio of *True Positives* to all the positives predicted by the model. Low precision: the more False positives the model predicts, the lower the precision.

c) Recall (Sensitivity):

Is the ratio of *True Positives* to all the positives in your Dataset. Low recall: the more False Negatives the model predicts, the lower the recall.

The idea of recall and precision seems to be abstract. Let me illustrate the difference in three real cases.

d) Accuracy:

Accuracy classification score. In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must exactly match the corresponding set of labels in y_true. Ground truth (correct) labels. Predicted labels, as returned by a classifier.

e) AUC ROC:

The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

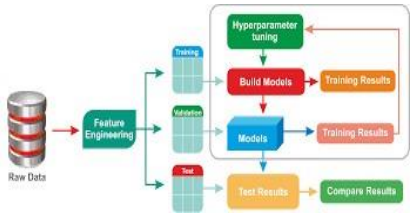
f) F1*:*

The F-score, also called the F1-score, is a measure of a model's accuracy on a dataset. ... The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

7.3. Hyper parameter tuning:

The process of selecting the appropriate hyperparameters for a learning algorithm is known as hyperparameter tuning. A hyperparameter is a model argument whose value is determined prior to the start of the learning process. Hyperparameter tweaking is the cornerstone to machine learning algorithms.

We may adjust or change the frequency of the model parameters using hyperparameters, which are similar to knobs on a radio. We can't change the model parameters directly; instead, we must change or set the hyperparameters. Hyperparameters are essentially algorithm properties.



- n_estimators = number of trees in the forest
- max_features = max number of features considered for splitting a node
- max_depth = max number of levels in each decision tree
- min_samples_split = min number of data points placed in a node before the node is split
- min_samples_leaf = min number of data points allowed in a leaf node
- bootstrap = method for sampling data points (with or without replacement)

Above are some examples of hyperparameters we use in Tree based algorithms. We can add cross validation as well in hyperparameters.

A) GridSearchCV:

The machine learning model is assessed for a range of hyperparameter values using the GridSearchCV technique. GridSearchCV is the name given to this method since it searches through a grid of hyperparameter values to find the optimal set of hyperparameters.

B) RandomizedSearchCV:

Because it only runs through a predetermined number of hyperparameter settings, RandomizedSearchCV overcomes the shortcomings of GridSearchCV. It moves randomly throughout the grid to discover the optimal collection of hyperparameters. This method eliminates the need for excessive computation.

8. Conclusion:

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 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%.

So, the accuracy of our best model is 90% which is by XGBoost so, Xgboost is our optimal model.

References-

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya
4. Towards DataScience
5. Data Science from scratch
6. Neptune
7. Builtin
8. Plosone 9. ijert