CUSTOMER RETENTION CASE STUDY

Submitted by:

ASHIKA YASMEEN

ACKNOWLEDGMENT

It gives me a great pleasure in presenting the project report on Customer Retention Using Machine Learning.

Sources:

Google.com

Wikipedia

INTRODUCTION

Customer satisfaction has been emerged as 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 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 e-commerce store have been identified as: service quality, system quality, information quality, trust and net benefit. The research furthermore investigated the factors that influence the online customers repeat purchase intention. The combination of both utilitarian value and hedonistic values are needed to affect the repeat purchase intention (loyalty) positively. The data is collected from the Indian online shoppers. Results indicate the e-retail success factors, which are very much critical for customer satisfaction.

Business Problem:

Customer retention also known as customer attrition or customer churn, is the loss of customers. It is an important and challenging problem for e-commerce and online businesses. Customer retention is where the customers have decided to end their relationship with a company, because they can cause a direct loss of revenue.

For any business in a designated period of time, customers can fall into 3 categories: Newly acquired customers, existing customers and churned customers. It costs more to acquire new customers than to retain existing ones. Because it costs more to acquire them, it is wise to work hard towards retaining them.

Background:

Machine learning is a subfield of Artificial Intelligence (AI) that works with algorithms and technologies to extract useful information from data. Machine learning methods are appropriate in big data since attempting to manually solve big data would be difficult without the support of machines. Machine learning attempts to solve problems algorithmically rather than purely mathematically. Therefore, it is based on creating algorithms that permit the machine to learn. Machine learning is categorized into two groups, which are supervised and unsupervised.

Supervised machine learning is an approach that is defined by its use of labeled datasets. These datasets are designed to train algorithms into classifying data or predicting outcomes accurately. Supervised learning can be separated into two types, classification and regression.

Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns in data without any human intervention. Unsupervised machine learning models are used for three main tasks, clustering, association and dimensionality reduction.

The performance will be measured upon predicting which of the Indian retailer is most recommended to friends. Since the prediction in many classification algorithms relies not only on a specific feature but on an unknown number of attributes that result in the value to be predicted. E-commerce has different number of features that may result in customer retention.

The data used in this project will be handled by using a combination of pre-processing methods to improve the predictions accuracy.

Literature:

Literature survey is the most important in any kind of research. Most of the literature study is based on articles from websites. This literature study attempts to construct a model based on classification techniques and how it can be applied to predict house price.

The literature study gives an overview of the articles that are related to this study, the feature engineering methods and the evaluation metrics that are used to measure the performance of the algorithms.

This paper presents how to make optimal use of Logistic Regression, Decision Tree Classifier and Random Forest Classifier. The system will give effective result on knowing the causes for customer retention.

Motivation:

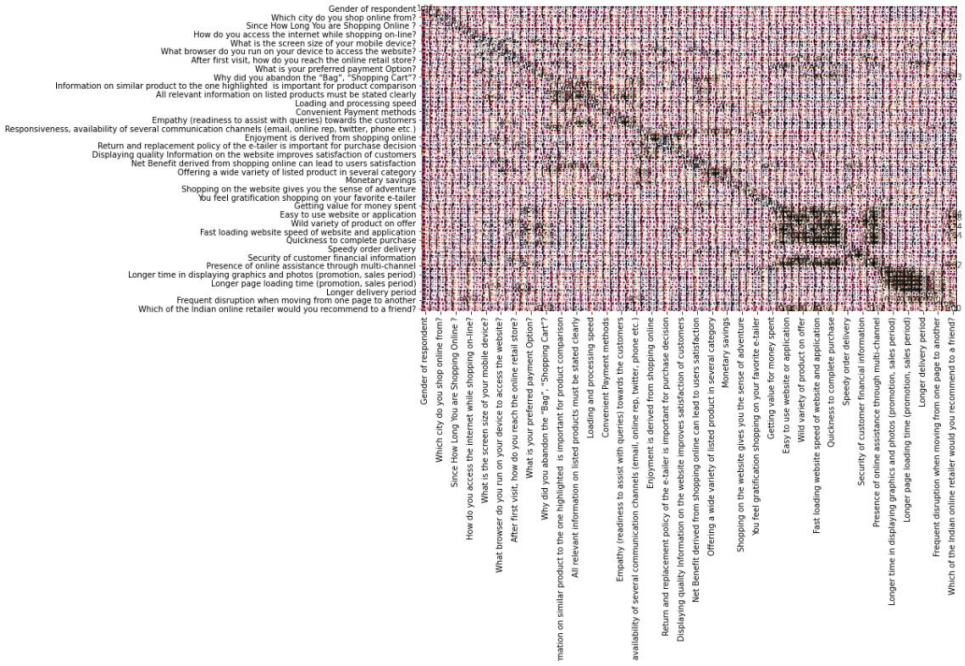
Every customer is different. So, there are several and completely different reasons for customer retention. Among them, we can find a lack of usage of the product, poor service or better pricing in other similar services. Hence, this project is developed to reduce customer retention by learning customer behaviour towards various changing factors in Indian e-commerce.

ANALYTICAL PROBLEM FRAMING

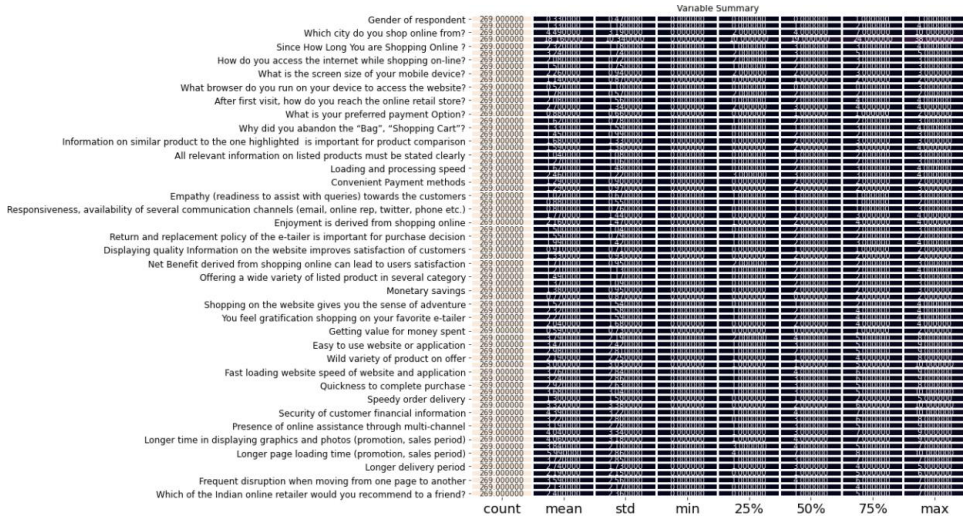
Mathematical/Analytical Modelling:

This experiment is done to pre-process the data and evaluate the prediction accuracy of the models. The experiment has multiple stages that are required to get the prediction results. These stages can be defined as:

- Pre-Processing: The dataset will be checked and pre-processed using certain methods. These methods have various ways of handling data. Thus, the pre-processing is done on multiple iterations where each time the accuracy will be evaluated with the used combination.
- Data splitting: Dividing the dataset into two parts to train the model with one and use the other in the testing. The dataset will be split 75% for training and 25% for testing.
- Evaluation: The accuracy of both datasets will be evaluated by measuring the accuracy score, confusion matrix and classification report when training the model with an evaluation of the on the test dataset with that are being predicted by the model.
- Correlation: Correlation analysis defines the strength of a relationship between two variables, which can be between two independent variables or one independent and one dependent variable. Correlation between the available features and output will be evaluated to identify whether the features have a negative, positive or zero correlation with the output variable.



- Describe of the dataset: Describe of the dataset is plotted using the heatmap which shows us the mean, standard deviation, maximum and minimum value of each column in the given dataset.



- Evaluation Metrics: The prediction accuracy will be evaluated by measuring the accuracy score, confusion matrix and classification report. Accuracy score is the ratio of number of correct predictions to the total number of input samples. It works well only if there are equal number of samples belonging to each class. A confusion matrix is an NxN matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. A classification report is a performance evaluation metric in machine

learning. It is used to show the precision, recall, F1 score and support of our trained classification model.

- Outliers are datapoints that are far away from other similar points. They may be due to variability in the measurement or may be due to experimental errors. Outliers should be excluded from the dataset in order to get the efficient and accurate prediction.

Methods to remove outliers:

1. Z-score: Call scipy.stats.zscore() with the given data-frame as its argument to get a numpy array containing the z-score of each value in a data-frame. Call numpy.abs with the previous result to convert each element in the data-frame to its absolute value.
2. Interquartile (IQR) range: It can be used to remove outliers present in the dataframe. IQR can be calculated using scipy.stats.iqr module.

- Skewness: Skewness is a measure of the asymmetry of the probability distribution of a real valued random variable about its mean.

Skewness:

```
cust_en.skew()
```

```
Gender of respondent                                                 0.741028
How old are you?                                                     0.680987
Which city do you shop online from?                                  0.313729
What is the Pin Code of where you shop online from?                  0.114537
Since How Long You are Shopping Online ?                            -0.276968
                                                                        ...
Longer delivery period                                              -0.147702
Change in website/Application design                                 0.354163
Frequent disruption when moving from one page to another            -0.100608
Website is as efficient as before                                    0.662084
Which of the Indian online retailer would you recommend to a friend?    0.583614
Length: 71, dtype: float64
```

Data Description:

In Machine Learning, the training data set is the actual dataset used to train the model for predicting the customer retention.

There are 269 rows and 71 columns

The columns consist of details of the customers like gender, age, city from where they are shopping online and their respective pin codes, from how long they are shopping online, how many times they have made online shopping past 1 year, access to the internet while shopping online, device used while shopping online, screen size of the device used, operating system of the device, from which browser online shopping is done, channel used while shopping online for the

first time, exploring done before doing online shopping for the first time, preferred payment option, how frequently did they abandon shopping and the reason for it, understanding of the content of the products, information on similar product, complete and relevant information on listed products, ease of navigation, processing speed, how user friendly is the website interface, convenient payment methods, trust, empathy, guarantee, responsiveness, availability of communication channels, information about monetary benefits and discounts, enjoyment from online shopping, convenience and flexibility, return and replacement policy, gaining access to loyalty programs, displaying quality information, user satisfaction, net benefit, listing several categories, quickness of purchase, availability of payment options, online assistance, delivery details, efficiency of websites, disruption while moving from one page to another, which retailer is recommended to a friend.

Data Pre-Processing:

Data pre-processing is an important step in machine learning to get highly accurate and reliable result. Data pre-processing helps in increasing the quality of data by filling in missing data's(NaN values), removing outliers, scaling the data.

There are many steps involved in data pre-processing:

- Data Cleaning helps to impute the missed values and removing outliers from the dataset.
- Data Integration integrates data from multiple sources into single dataset.
- Data Transformation such as normalization helps in improving the accuracy and efficiency of algorithms involved in machine learning.
- Data Reduction reduces the data size by dropping out redundant features using feature selection and feature extraction techniques.

Treating null values

Sometimes there can be certain columns which may contain the null values used to indicate the missing or unknown values. In our dataset there are no null values present.

Converting labels into numeric

In machine learning, we usually deal with datasets which contain multiple labels in one or more column. These labels can be in the form of alphabets or numbers. To make the data understandable or in human readable form, the training data is often labelled in words.

In our dataset all the columns have categorical values. These columns are converted using Label Encoder.

Label Encoder refers to converting the labels into numeric form so as to convert it into the machine-readable form. It is an important step in data pre-processing.

Label encoding in python can be imported from Sklearn library. Sklearn provides a very efficient tool for encoding.

Label Encoding:

```
from sklearn.preprocessing import LabelEncoder

cust_en=pd.DataFrame()
le=LabelEncoder()
for column in cust.columns:
    cust_en[column]=le.fit_transform(cust[column])
cust_en
```

Input-Output Relationship:

The dependent variable or target or output depends on the features or input variables given in the dataset.

Tools Used:

Hardware: The needed time to train the model depends on the capability of the used system during the experiment. Some libraries use GPU resources over the CPU to take a shorter time to train a model.

Operating System	Windows 10
Processor	CORE i3
RAM	16GB

Language: Python

- Python is widely used in numeric and scientific computing.
- Scipy is a collection of packages for mathematics, science and engineering.
- Pandas is a data analysis and modelling library.

Libraries:

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Scikit Learn

MODEL DEVELOPMENT AND EVALUATION

Model:

Classification model is used. Classification refers to a predictive modeling problem where a class label is predicted for a given set of input data. It is a supervised technique. Regression models a target prediction based on independent variables.

Logistic Regression:

Linear regression is a

learning algorithm based on supervised learning. Logistic regression is a classification algorithm, used when the value of the target variable is categorical in nature. It is most commonly used when the data has binary input, so when it belongs to one class or the other, or is either a 0 or 1.

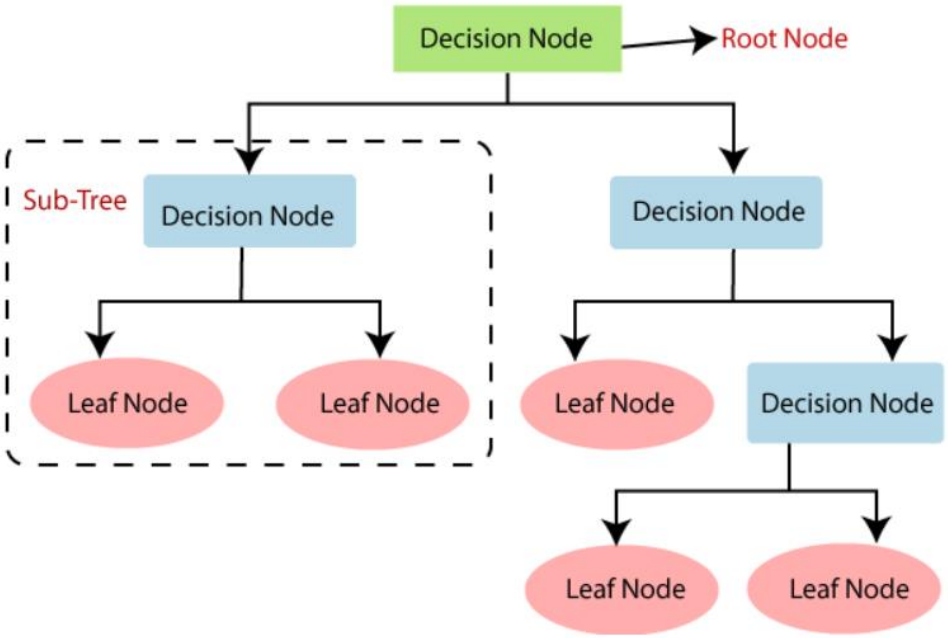
Decision Tree Classifier:

It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a decision tree, there are two nodes, which are the decision node and leaf node. Decision nodes are used to make any decision and have multiple

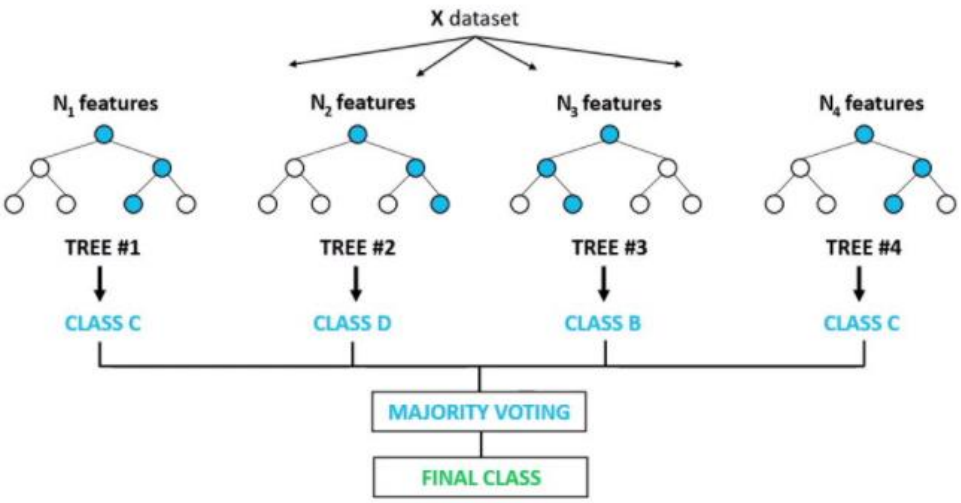
branches, whereas leaf nodes are the output of those decisions and do not contain any further branches.

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like-structure.



Random Forest Classifier:

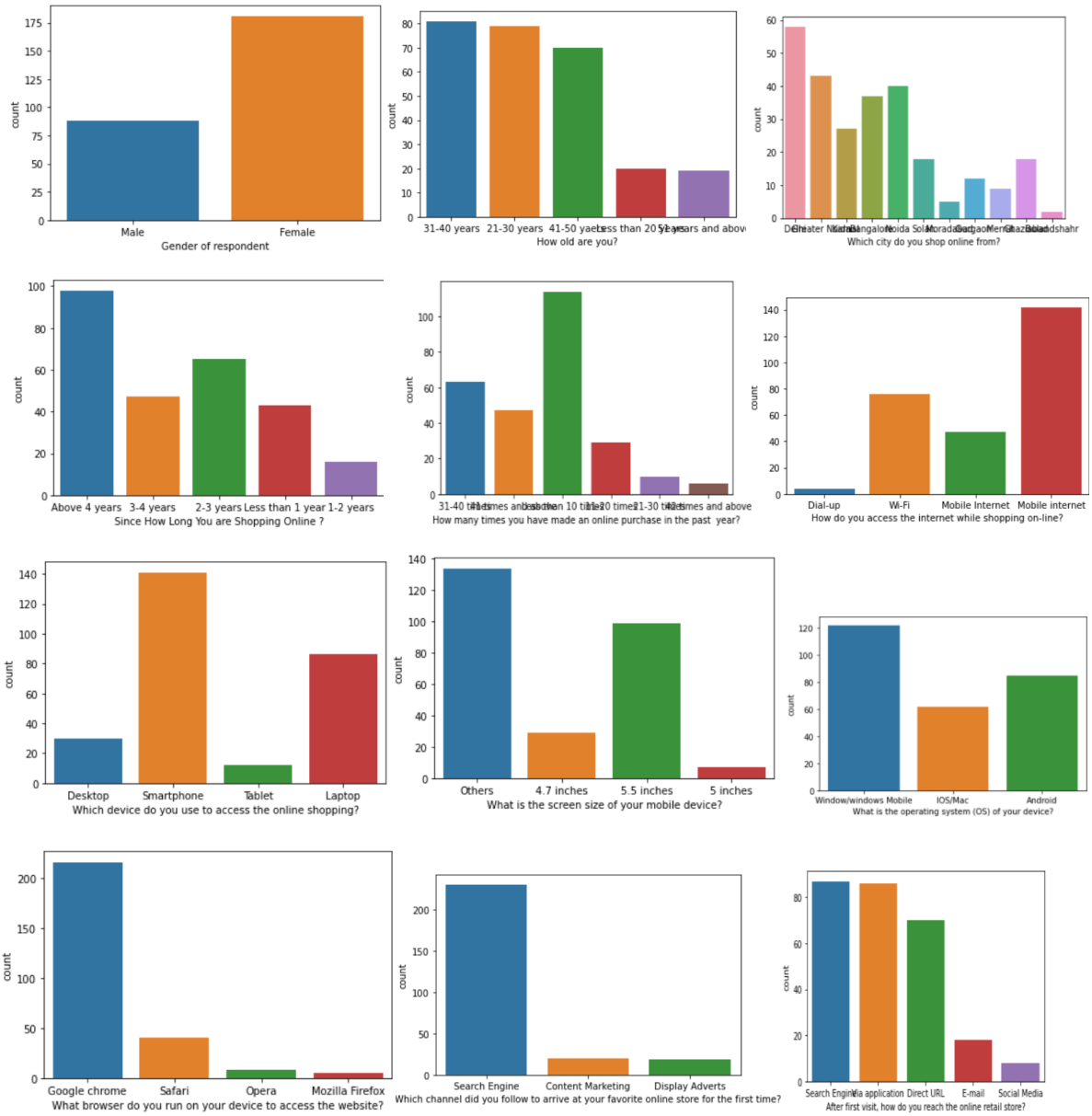
It is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each individual tree and try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

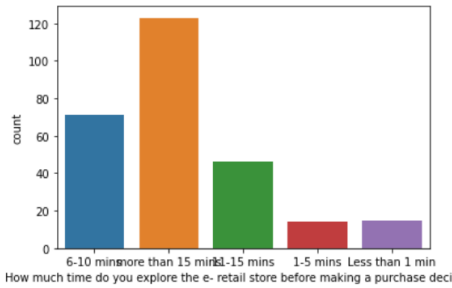


Algorithms Used:

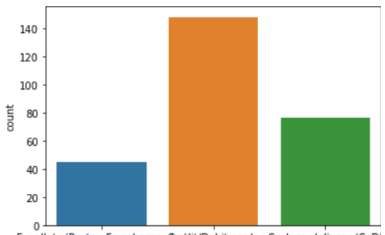
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier

Visualizations:

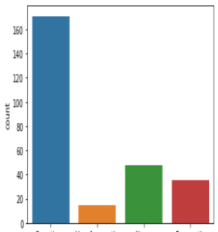




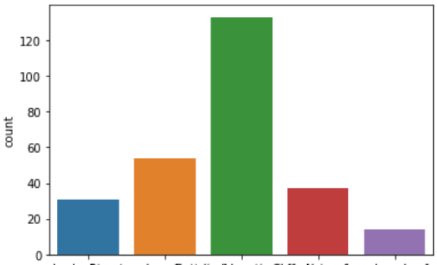
How much time do you explore the e- retail store before making a purchase decision?



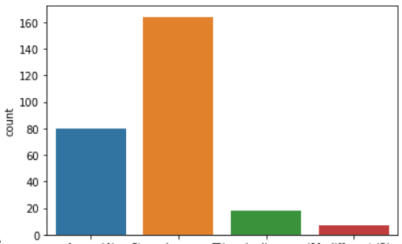
What is your preferred payment Option?



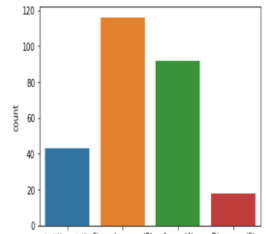
How frequently do you abandon (selecting an items and leaving without making payment) your shopping cart



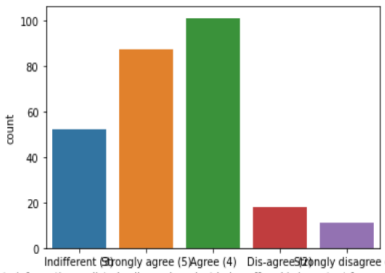
Why did you abandon the “Bag”, “Shopping Cart”?



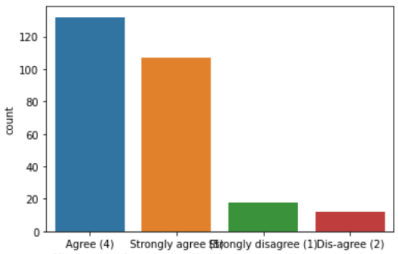
The content on the website must be easy to read and understand



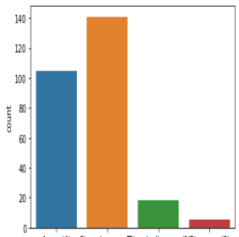
Information on similar product to the one highlighted is important for product comparison



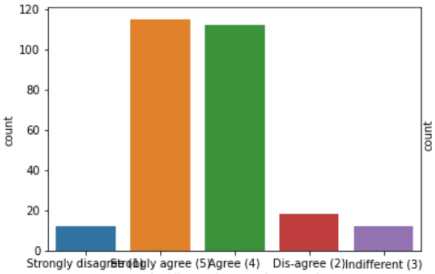
Complete information on listed seller and product being offered is important for purchase decision.



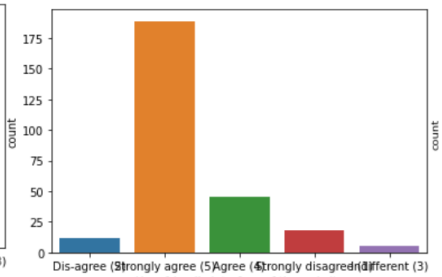
All relevant information on listed products must be stated clearly



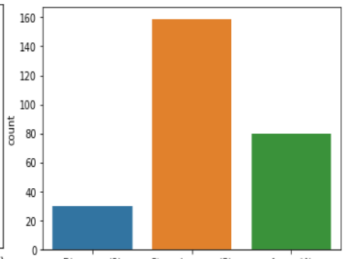
Ease of navigation in website



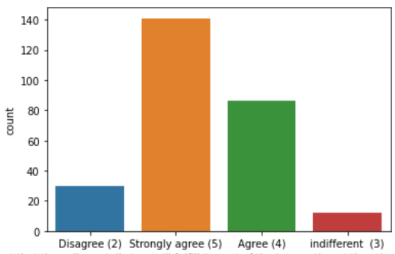
Loading and processing speed



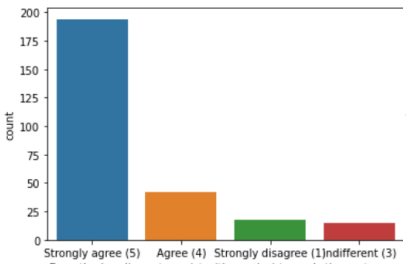
User friendly Interface of the website



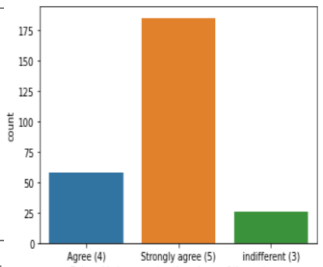
Convenient Payment methods



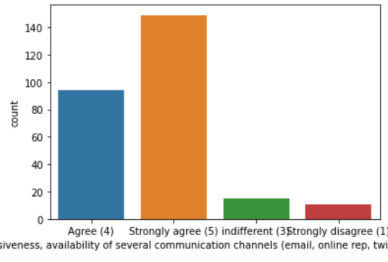
Trust that the online retail store will fulfill its part of the transaction at the stipulated time



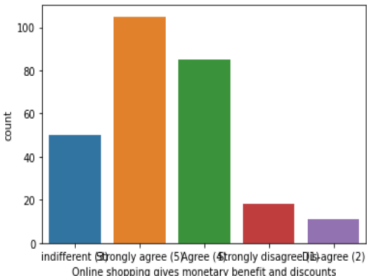
Empathy (readiness to assist with queries) towards the customers



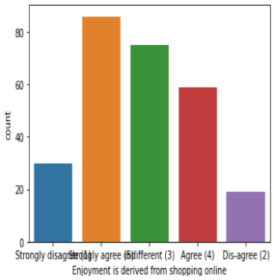
Being able to guarantee the privacy of the customer



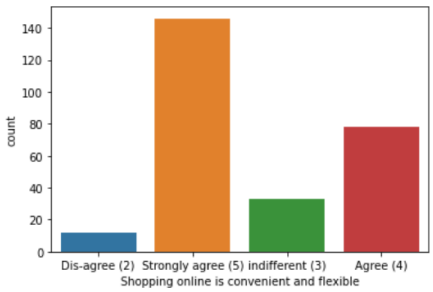
Responsiveness, availability of several communication channels (email, online rep, twitter, phone etc.)



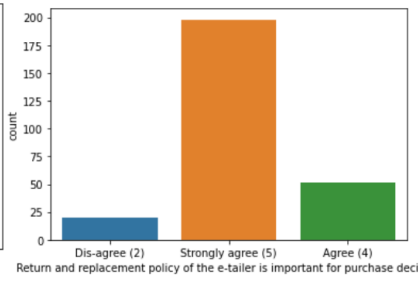
Online shopping gives monetary benefit and discounts



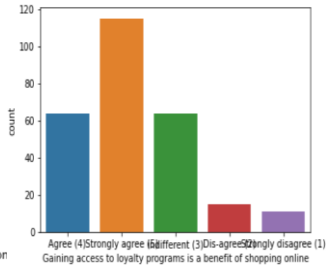
Enjoyment is derived from shopping online



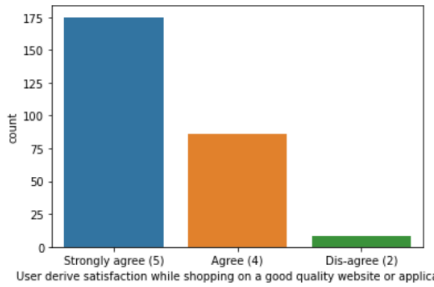
Shopping online is convenient and flexible



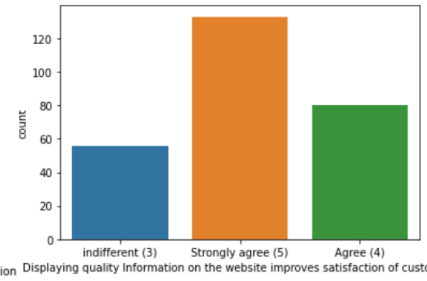
Return and replacement policy of the e-tailer is important for purchase decision



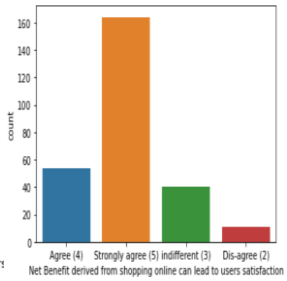
Gaining access to loyalty programs is a benefit of shopping online



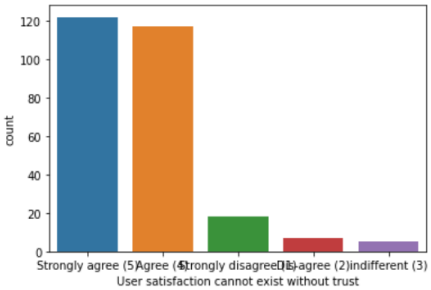
User derive satisfaction while shopping on a good quality website or application



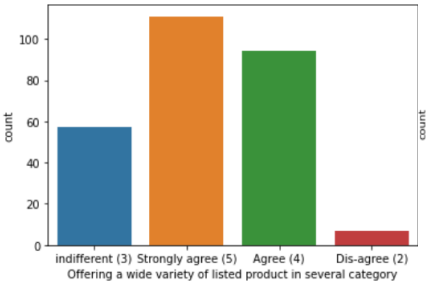
Displaying quality Information on the website improves satisfaction of customers



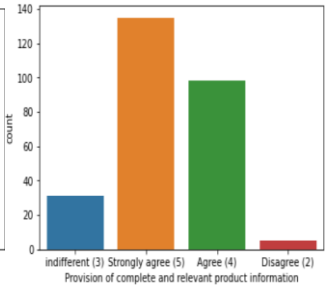
Net Benefit derived from shopping online can lead to users satisfaction



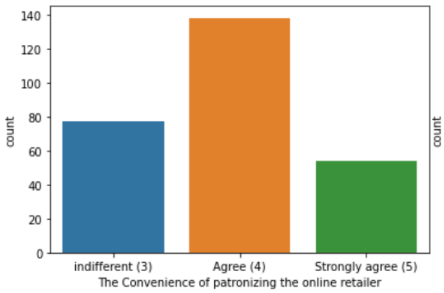
User satisfaction cannot exist without trust



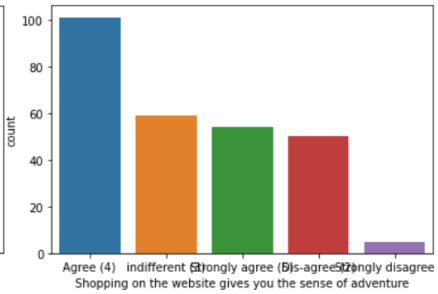
Offering a wide variety of listed product in several category



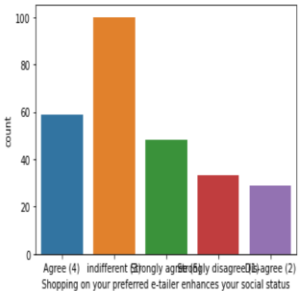
Provision of complete and relevant product information



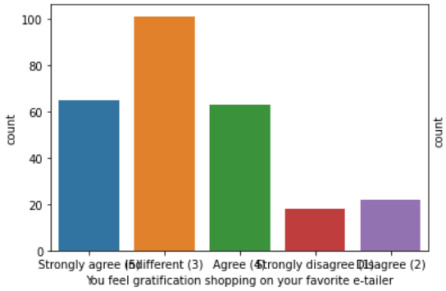
The Convenience of patronizing the online retailer



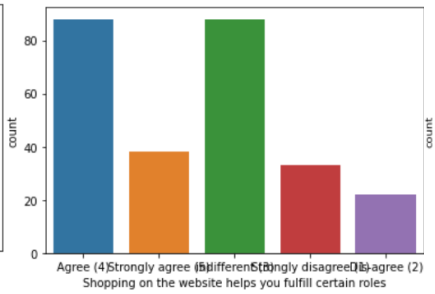
Shopping on the website gives you the sense of adventure



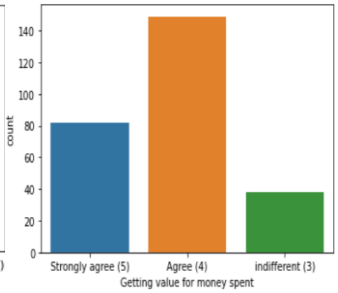
Shopping on your preferred e-tailer enhances your social status



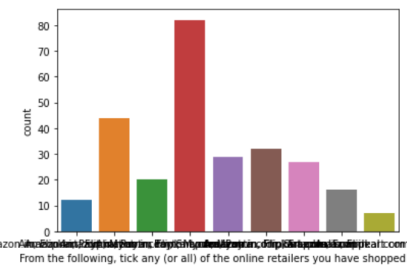
You feel gratification shopping on your favorite e-tailer



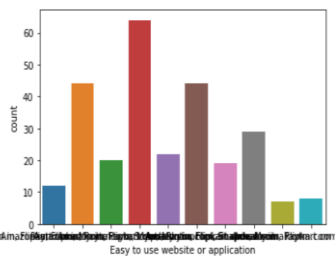
Shopping on the website helps you fulfill certain roles



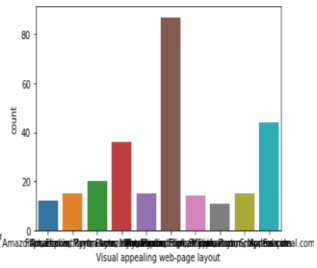
Getting value for money spent



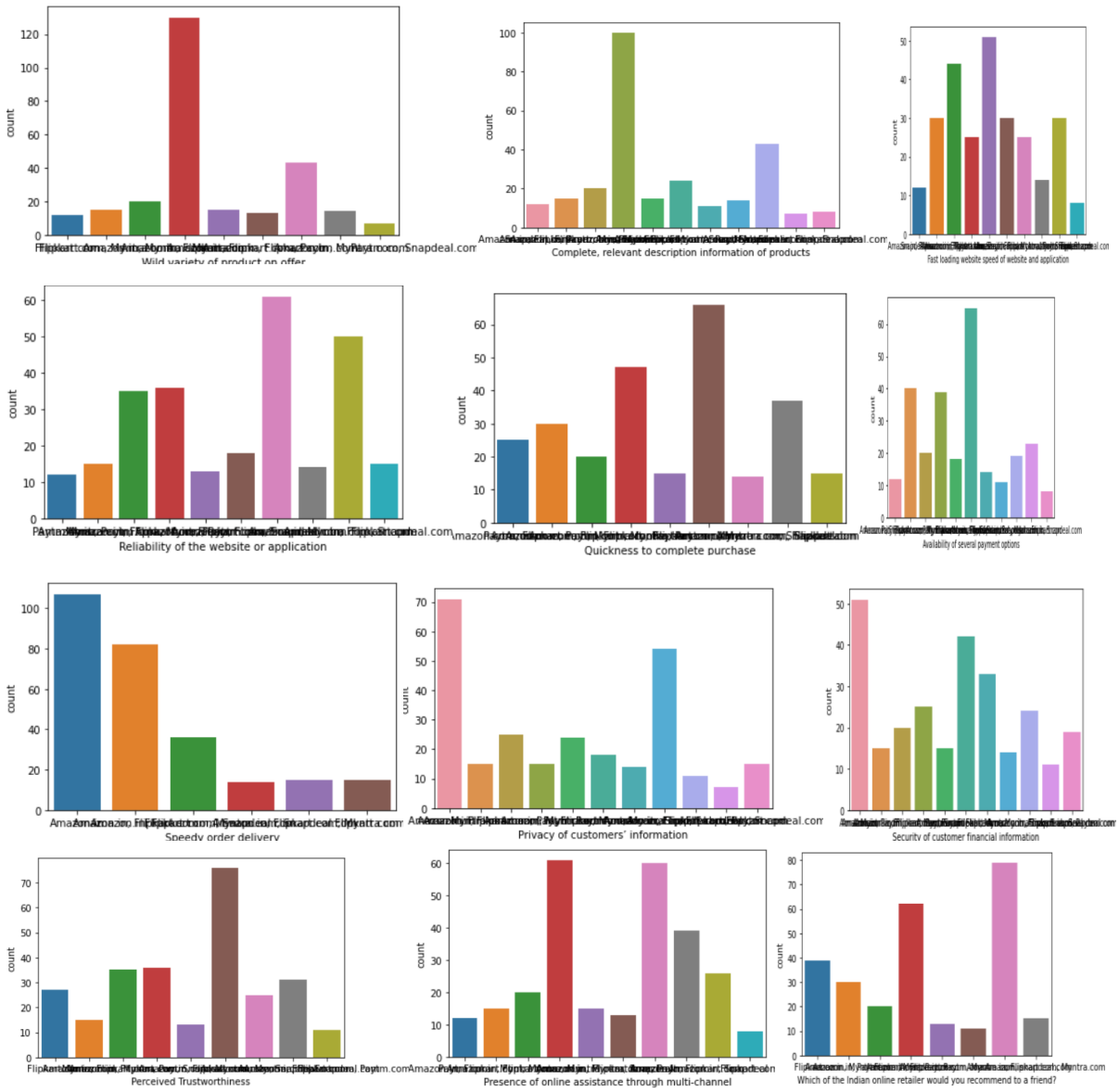
From the following, tick any (or all) of the online retailers you have shopped from;



Easy to use website or application



Visual appealing web-page layout



Interpretation of Results:

Many machine learning algorithms are used to predict. Using these algorithms is beneficial so that the result can be as near to the claimed results. However, the prediction accuracy of these algorithms depends heavily on the given data when training the model. If the data is in bad shape, the model will be overfitted and inefficient, which means that data pre-processing is an important part of this experiment and will affect the final results. Thus, multiple combinations of pre-processing methods need to be tested before getting the data ready to be used in training.

From Visualizations it is clear that how each feature from the dataset is relatable to the house price. In pre-processing, unnecessary feature is removed, outliers are removed and skewness from the input variables are

removed. The data's are scaled and transformed for better training purpose. However, we have got 100% accuracy with all the models taken here, the chosen algorithm is Random Forest Classifier.

CONCLUSION

The cost of the product, the reliability of the E-commerce company and the return policies play an equally important role in deciding the buying behaviour of online customers. The cost is an important factor as it was the basic criteria used by online retailers to attract customers. The reliability of the E-commerce company is also important, as it is even required in offline retail. It is important because customers are paying online., so they need to be sure of security of the online transaction. The return policies are important because in online retail customer does not get to feel the product. Thus, the customer wants to be sure that it will be possible to return the product if it is not liked in real. Whereas, the logistics factor, which included cash on delivery option, one day delivery and the quality of packaging plays a role in the process though these are important factors.

All the websites were not equally preferred by online customers. Amazon was the most preferred followed by Flipkart. These two companies are most trusted in the industry and have a huge reliability. Hence, these two companies have comparatively lesser possibilities of customer retention.

Limitations:

This study did not cover all the classification algorithms; instead, it is focused on the chosen algorithm, starting from the basic regression techniques to the advanced techniques.