

CUSTOMER RETENTION PROJECT

Submitted by:

Praveen Kumar Singh

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project. Various websites we used like *https://www.ngdata.com/, https://www.crazyegg.com/blog/customer-retention/, https://en.wikipedia.org/wiki/Customer\_retention* etc.

Presentation, inspiration and motivation have always played a key role in the success of any venture.

I express my sincere thanks to Dr. Deepika mam, Vishal Sir and Ram Sir of Datatrained Acedemy.

**INTRODUCTION**

* **Business Problem Framing**

Customer retention refers to the ability of a company or product to retain its customers over some specified period. High customer retention means customers of the product or business tend to return to, continue to buy or in some other way not defect to another product or business, or to non-use entirely. Selling organizations generally attempt to reduce [customer defections](https://en.wikipedia.org/wiki/Customer_switching). Customer retention starts with the first contact an organization has with a customer and continues throughout the entire lifetime of a relationship and successful retention efforts take this entire lifecycle into account. A company's ability to attract and retain new customers is related not only to its [product](https://en.wikipedia.org/wiki/Product_(business)) or services, but also to the way it services its existing customers, the value the customers actually perceive as a result of utilizing the solutions, and the reputation it creates within and across the [marketplace](https://en.wikipedia.org/wiki/Marketplace).

* **Conceptual Background of the Domain Problem**

E-commerce is growing rapidly since its first day. The credit card company reported the internet purchase of their customers has researched 13 billion dollar mark in the year of 2000. They still expected to have a tremendous grows in the future. The online transaction will be one if the major activity of credit card companies. In some areas the E-commerce has already become the most popular format of business, such as books, CDs and magazines. According to David (2000) Amazon now selling millions books, CDs, and DVDs to more than 6.2 million customers in more than 160 different countries and areas. (David, 2000).

With the exceptional rapid growth of the E-commerce business also come the great challenges. Customer relationship management in the E-commerce area is one of the areas require deeper research. In the 80‟s the marketing approach started to focus on customers than products. The concept of relationship marketing became the major marketing approach. Relationship marketing put more attention on attracting, maintaining and enhancing customer relationship. With the technology involving business combine the information technology and relationship marketing together to improve the efficiency of relationship marketing strategies, it forms the CRM concept. The ECRM concept is an involvement from CRM. Because the web technology and internet technology are becoming more and more wildly used and these technologies could give the businesses more durability to stay with their customer information, CRM concept adopt these technology to provide the better supporting for businesses- ECRM. 2 Since the adoption of relationship marketing lots of the research and data from the real businesses show that to retain an existing customer is far more economic than to require a new one. Business starts to invest more and more resource and efforts into this area. At the same time academic researchers also start to investigate different merits that enhance the customer retention. Also the E-commerce environment also is trying to catch the same trends because of the unique characters of this industry. Because of the unique characters of this industry, E-commerce business cannot just adopt the classic theories from relationship marketing. These businesses need some theories in the E-commerce context to support their operation.

* **Review of Literature**

Customer acquisition, customer retention and customer development are the three elements that form a customer lifecycle. A customer is acquired into the company as a part of reaching the ambition to grow the business. Acquired customers form the base of customer retention – without any customers, there is no churn to prevent or value to enhance. (Buttle, 2009.)

There is a great amount of literature about customer loyalty, customer satisfaction etc. But not many of researches are focus on the specific topic of customer retention. Customer retention should be a bigger topic. And also there are quite many different characters in the context of E-commerce compare to the regular businesses. So this is the motivation and purpose of this thesis is to gain better understanding of how Ecommerce business should retain their customer.

As customer-perceived value has a positive and significant effect on customer satisfaction, insurance service companies must focus on improving service quality and company image to achieve customer satisfaction, which would in turn yield customer loyalty (Nguyen et al., [2018](https://www.tandfonline.com/doi/full/10.1080/23311975.2020.1738200)). Adiati and Dinna ([2014](https://www.tandfonline.com/doi/full/10.1080/23311975.2020.1738200)) similarly concluded that, in the internet service provider industry, there is a positive effect of customer-perceived value on customer satisfaction. According to Mbango ([2019](https://www.tandfonline.com/doi/full/10.1080/23311975.2020.1738200)), customer value has an influence on customer satisfaction. Samudro et al. ([2020](https://www.tandfonline.com/doi/full/10.1080/23311975.2020.1738200)) as well as Susanti et al. ([2020](https://www.tandfonline.com/doi/full/10.1080/23311975.2020.1738200)) also confirmed that perceived value exerts a stronger influence on customer satisfaction in the chemical industry. Based on these empirical findings, we hypothesised:

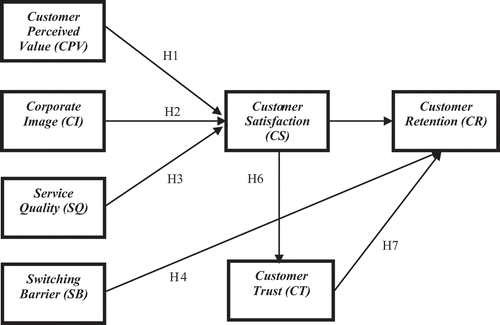
H1: Customer-perceived value has a significant effect on customer satisfaction.

Several previous studies have indicated that service quality significantly affects customer satisfaction. For example, Ashraf et al. ([2018](https://www.tandfonline.com/doi/full/10.1080/23311975.2020.1738200)) conducted a case study on five lines of business engaged in services, hotels, hospitals, education, and banks in Pakistan, reporting a significant effect of service quality on customer satisfaction. Furthermore, service quality has been found as an important determinant of customer satisfaction in retail banking (Caruana, [2002](https://www.tandfonline.com/doi/full/10.1080/23311975.2020.1738200)), the game industry (Wu, [2014](https://www.tandfonline.com/doi/full/10.1080/23311975.2020.1738200)), chemical industry (Susanti et al., [2020](https://www.tandfonline.com/doi/full/10.1080/23311975.2020.1738200)), international medical travel sector (Han & Hyun, [2015](https://www.tandfonline.com/doi/full/10.1080/23311975.2020.1738200)), and in cargo services (Arief et al., [2019](https://www.tandfonline.com/doi/full/10.1080/23311975.2020.1738200)). This has resulted in the following hypothesis:

H3: Service quality has a significant effect on customer satisfaction.

* **Motivation for the Problem Undertaken**

Based on the empirical studies reviewed above, it is hypothesised that customer-perceived value, corporate image, and service quality influence customer retention through customer satisfaction, while switching barriers have a direct effect on customer retention. It is further hypothesised that customer satisfaction affects customer retention through customer trust. The conceptual framework is presented



**Analytical Problem Framing**

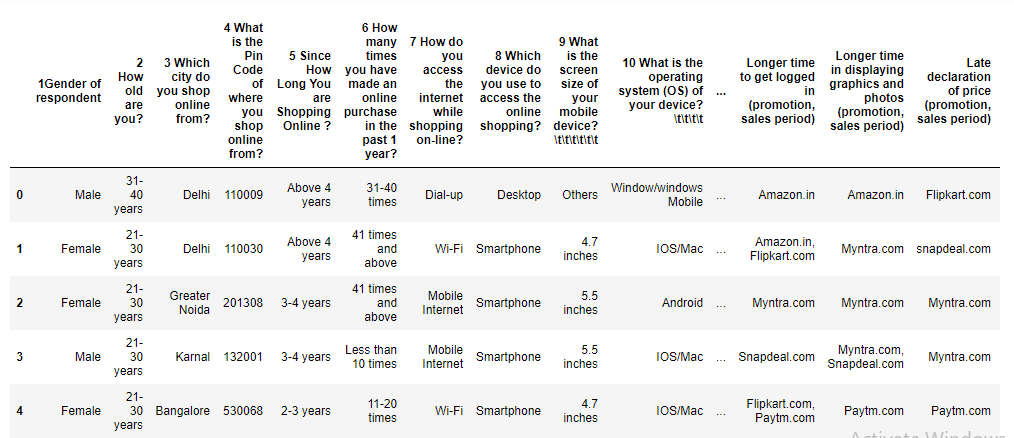
* **Mathematical/ Analytical Modeling of the Problem**

While creating the model, we are doing mathematical/analytical modelling of the problem. We generally do duplicates removal, null values removal, NaN values removal, getting statistical summary, replacing unwanted values with some useful values, checking the correlation of the attributes.

* **Data Sources and their formats**

Here data sources refer to data which we are going to use for model creation. Generally we get the data from any firm which wants to create a model for their need. We can get data in any format; say it will be in csv, xls, xlsx format etc. Sometimes it may happen that we can get data without column names also.

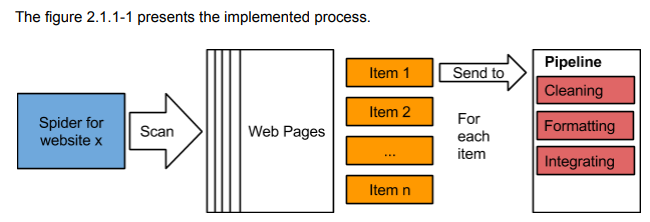
Below is the screenshot of the data.

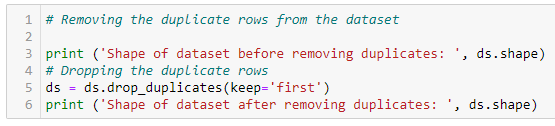


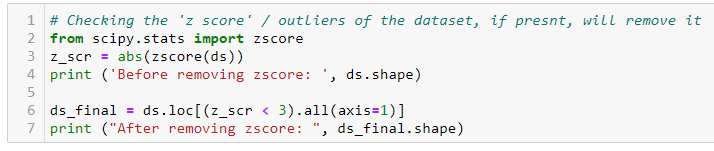
* **Data Preprocessing Done**

For creating the model, data should be neat and clean. For that we perform some pre-processing on the data. Pre-processing of data is very important step in model creation. At what extent our model performs well is all depends on pre-processing.

While reading the dataset file, at that time only, we consider some unwanted values as NaN values. We remove the duplicates, remove the null values, NaN values, remove the outliers, scale the data etc.







* **Data Inputs- Logic- Output Relationships**

Generally, we get the data in the form of file. Dataset contains multiple columns which plays an important role for estimating the output. If one of the input parameter varies output also varies. For example, for customer retention, there are many factors which affect it. The factors, we can say , ‘Which city do you shop online from’, ‘Gaining access to loyalty programs is a benefit of shopping online’, ‘Net Benefit derived from shopping online can lead to users satisfaction’, ‘Quickness to complete purchase’, ‘Frequent disruption when moving from one page to another’, ‘Late declaration of price (promotion, sales period)’, etc.

Satisfaction is “a judgment that a product or service feature, or the product or service itself, provides a pleasurable level of consumption-related fulfillment” [60]. Customer satisfaction is an effective state connected with positive feelings as a resultant of overall usage experience evaluation [12]. Customer satisfaction level is an outcome of an analysis by customer of how fine their expectations matched with experience but measurement of satisfaction has always remained a difficult question for researchers [56]. In the situations where service performance exceeds more than customer’s expectations then customer satisfaction is increased, this is positive confirmation.

* **State the set of assumptions (if any) related to the problem under consideration**

In this project, our aim is to predict the “*Which of the Indian online retailer would you recommend to a friend*?” There are multiple factors affecting the target.

Below are the factors which contribute in predicting the ‘Recommend retailer’.

*'1Gender of respondent', '2 How old are you? ',*

*'3 Which city do you shop online from?',*

*'4 What is the Pin Code of where you shop online from?',*

*'5 Since How Long You are Shopping Online ?',*

*'6 How many times you have made an online purchase in the past 1 year?',*

*'7 How do you access the internet while shopping on-line?',*

*'8 Which device do you use to access the online shopping?',*

*'9 What is the screen size of your mobile device?\t\t\t\t\t\t ',*

*'10 What is the operating system (OS) of your device?\t\t\t\t ',*

*'11 What browser do you run on your device to access the website?\t\t\t ',*

*'12 Which channel did you follow to arrive at your favorite online store for the first time’, '13 After first visit, how do you reach the online retail store?\t\t\t\t ',*

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

*'15 What is your preferred payment Option?\t\t\t\t\t ',*

*'16 How frequently do you abandon (selecting an items and leaving without making payment) your shopping cart?\t\t\t\t\t\t\t ',*

*'17 Why did you* *abandon the “Bag”, “Shopping Cart”?\t\t\t\t\t ',*

*'18 The content on the website must be easy to read and understand',*

*'19 Information on similar product to the one highlighted is important for product comparison',*

*'20 Complete information on listed seller and product being offered is important for purchase decision.',*

*'21 All relevant information on listed products must be stated clearly',*

*'22 Ease of navigation in website', '23 Loading and processing speed',*

*'24 User friendly Interface of the website',*

*'25 Convenient Payment methods',*

*'26 Trust that the online retail store will fulfill its part of the transaction at the stipulated time',*

*'27 Empathy (readiness to assist with queries) towards the customers',*

*'28 Being able to guarantee the privacy of the customer',*

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

*'30 Online shopping gives monetary benefit and discounts',*

*'31 Enjoyment is derived from shopping online',*

*'32 Shopping online is convenient and flexible',*

*'33 Return and replacement policy of the e-tailer is important for purchase decision',*

*'34 Gaining access to loyalty programs is a benefit of shopping online',*

*'35 Displaying quality Information on the website improves satisfaction of customers',*

*'36 User derive satisfaction while shopping on a good quality website or application',*

*'37 Net Benefit derived from shopping online can lead to users satisfaction',*

*'38 User satisfaction cannot exist without trust',*

*'39 Offering a wide variety of listed product in several category',*

*'40 Provision of complete and relevant product information',*

*'41 Monetary savings',*

*'42 The Convenience of patronizing the online retailer',*

*'43 Shopping on the website gives you the sense of adventure',*

*'44 Shopping on your preferred e-tailer enhances your social status',*

*'45 You feel gratification shopping on your favorite e-tailer',*

*'46 Shopping on the website helps you fulfill certain roles',*

*'47 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', 'Wild variety of product on offer',*

*'Complete, relevant description information of products',*

*'Fast loading website speed of website and application',*

*'Reliability of the website or application',*

*'Quickness to complete purchase',*

*'Availability of several payment options', 'Speedy order delivery ',*

*'Privacy of customers’ information',*

*'Security of customer financial information',*

*'Perceived Trustworthiness',*

*'Presence of online assistance through multi-channel',*

*'Longer time to get logged in (promotion, sales period)',*

*'Longer time in displaying graphics and photos (promotion, sales period)',*

*'Late declaration of price (promotion, sales period)',*

*'Longer page loading time (promotion, sales period)',*

*'Limited mode of payment on most products (promotion, sales period)',*

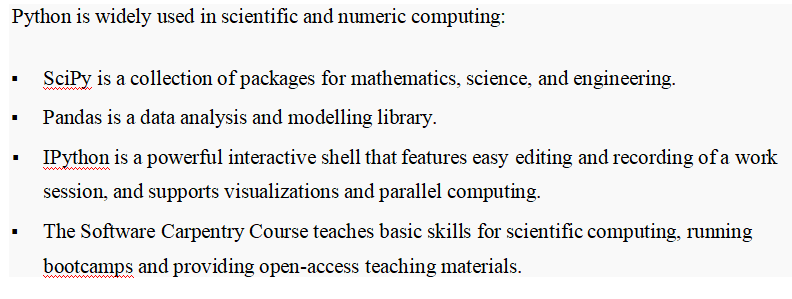
*'Longer delivery period', 'Change in website/Application design',*

*'Frequent disruption when moving from one page to another',*

*'Website is as efficient as before',*

*'Which of the Indian online retailer would you recommend to a friend*

* **Hardware and Software Requirements and Tools Used**



**Libraries Used for this Project include –**

* *Pandas*
* *NumPy*
* *Matplotlib*
* *Seaborn*
* *LineraRegression*
* *Seaborn*
* *Sklearn*

Various modules used from sklearn like linear\_model, model\_selection, neighbors, GradientBoostingRegressor, DecisionTreeRegressor etc.

**Model/s Development and Evaluation**

* **Identification of possible problem-solving approaches (methods)**

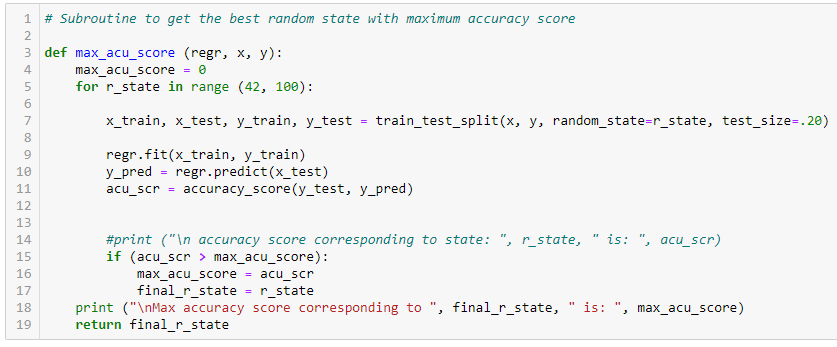
Model creation has several approaches. We are using statistical method for getting description of the dataset. Corr() method is used to check the correlation of the dataset. We are replacing unwanted values with meaningful values. StandardScalar is used to scale the data. We used the unique method to check the unique values of columns.

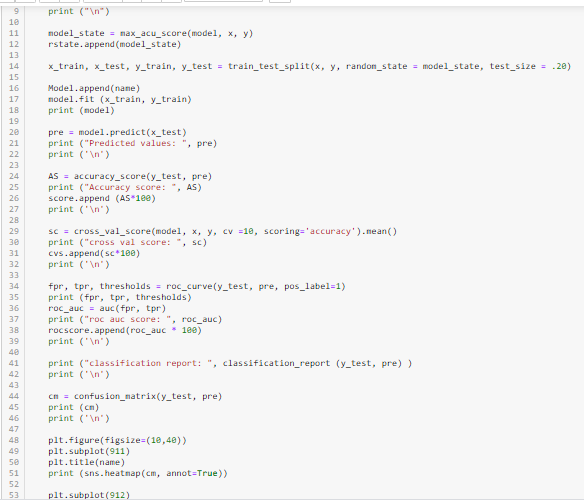
* **Testing of Identified Approaches (Algorithms)**

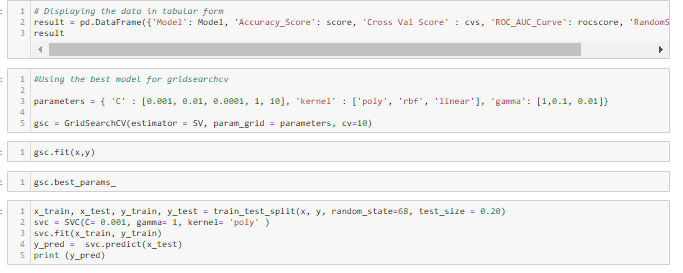
Below algorithms are used to train and test the dataset.

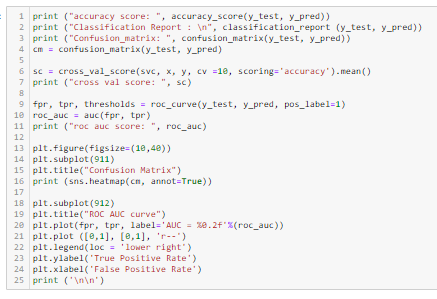
* KNeighborsClassifier
* GridSearchCV
* SVC
* DecisionTreeClassifier
* RandomForestClassifier
* **Run and Evaluate selected models**

All the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.









* **Key Metrics for success in solving problem under consideration**

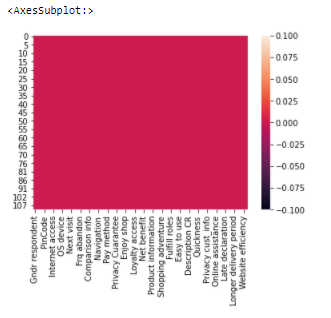
Key metrics used to get the best model. It includes clean and proper data. Data should be validated before proceeding further so we won’t get stuck in creating the model. We should clean the garbage data. Various plots are plotted to get the idea of how data vary, how the variation of one data affects other. We can check if there is positive or negative relation in data.

* **Visualizations**

For visualization, we used *seaborn* and *matplotlib.pyplot* libraries. We used various plots as mentioned below to get the idea of how data varies with each other.

1. **Heatmap for checking null values**

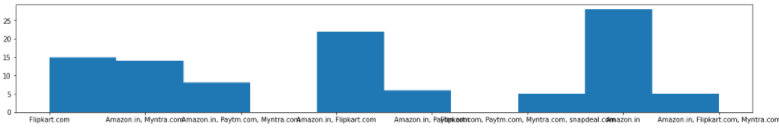
Below heatmap is used to check the null values in the dataset.

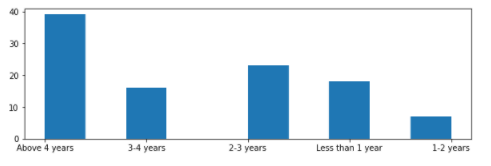


It shows that there are no null values in the dataset.

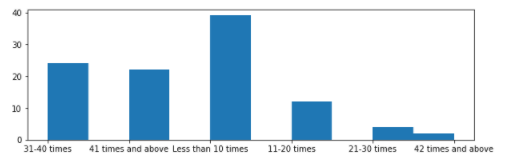
1. **Histograms**

A histogram is the most commonly used graph to show frequency distributions.

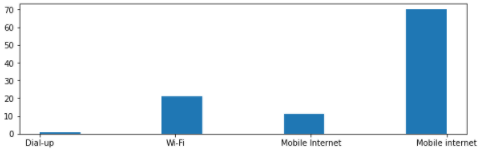




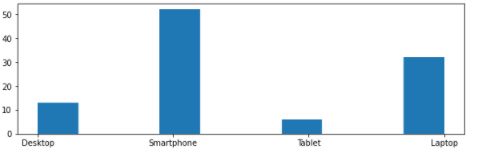
It shows that most of the shopping period is of above 4 years.



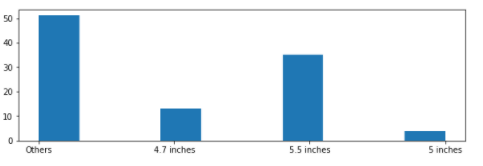
Many of the customers purchased less than 10 times it seems.



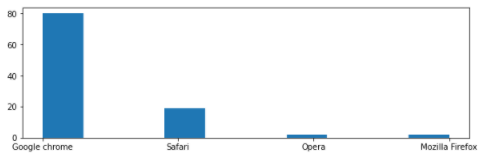
It seems many customers uses mobile internet for e-tailing.



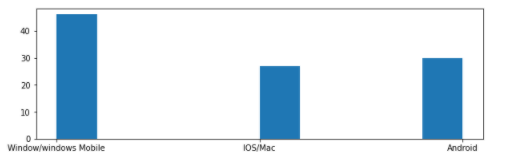
It shows many customers are using smartphones for shopping.



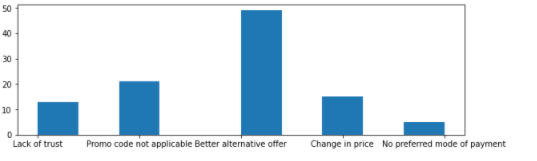
Many customers device screen size is other than 4.7, 5.5 and 5 inches.

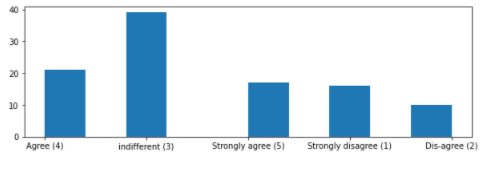


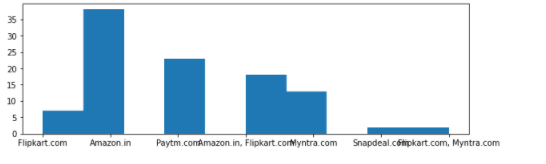
Many of the customers are using Google chrome for shopping.

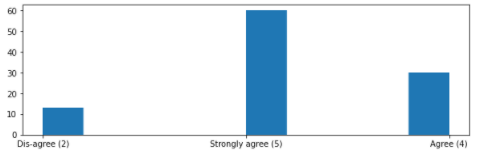


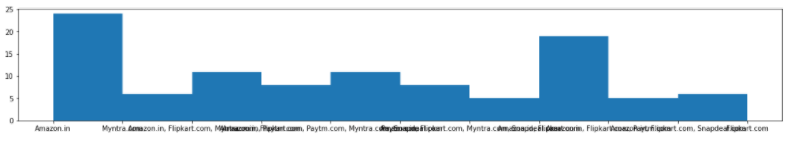
Many customers are using ‘Windows’ mobile.







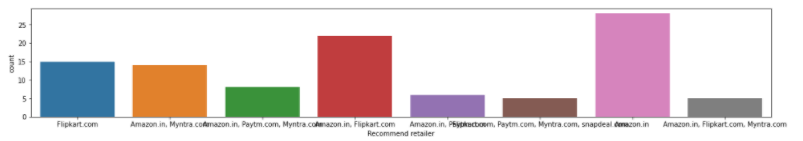




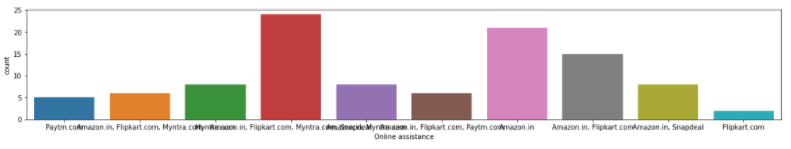
It seems Amazon has good customer privacy policy.

1. **Countplot**

A count plot can be thought of as a histogram across a categorical, instead of quantitative, variable. The basic API and options are identical to those for [barplot()](https://seaborn.pydata.org/generated/seaborn.barplot.html" \l "seaborn.barplot" \o "seaborn.barplot), so you can compare counts across nested variables.

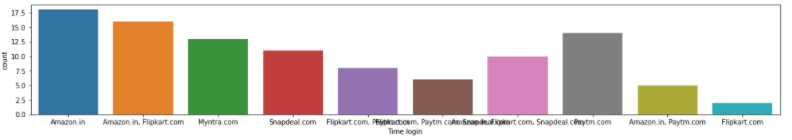


It seems most recommended retailer is Amazon.com

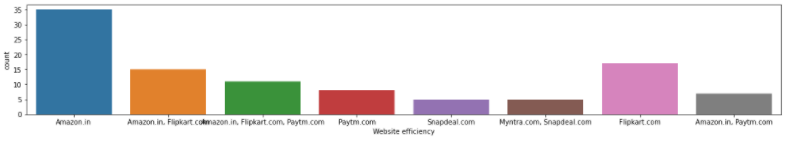


It seems Flipkart amd Myntra has good customer assistance.





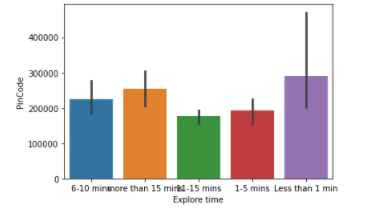
It seems Longer time to get logged in (promotion, sales period) is Amazon.



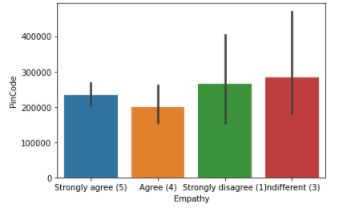
It looks Amazon has good Website efficiency.

1. **Barplot**

A bar chart is used when you want to show a distribution of data points or perform a comparison of metric values across different subgroups of your data.

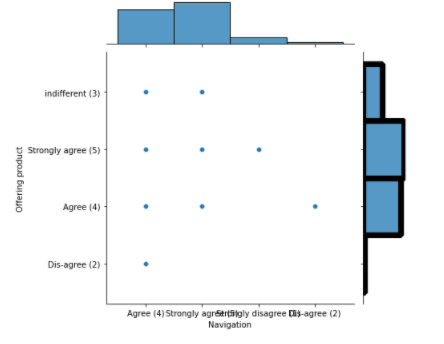
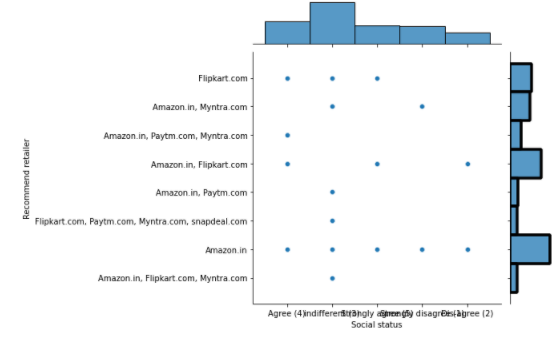


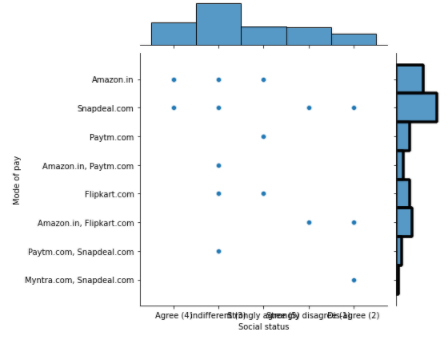
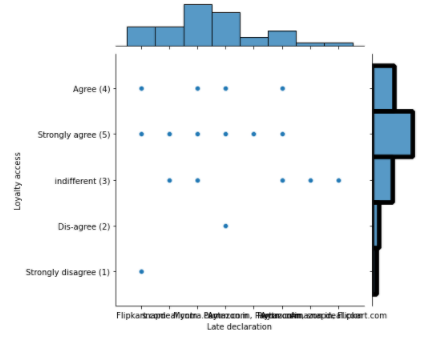
Pincode 30000 has more explore time it seems.



1. **Jointplot**

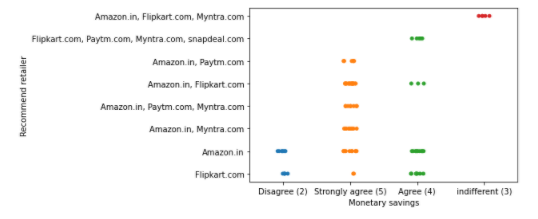
jointplot displays a relationship between 2 variables (bivariate) as well as 1D profiles (univariate) in the margins.

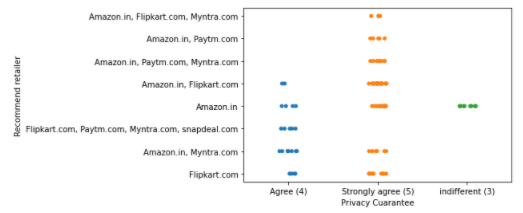
 

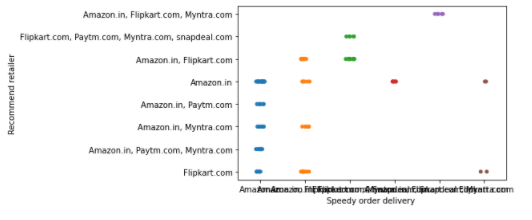
 

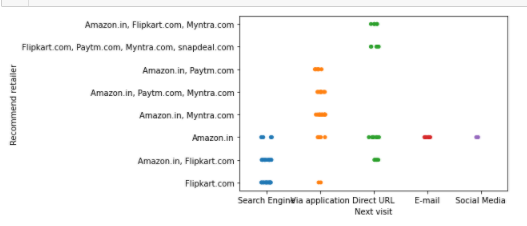
1. **StripPlot**

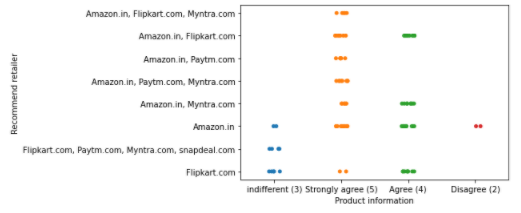
A strip plot can be drawn on its own, but it is also a good complement to a box or violin plot in cases where you want to show all observations along with some representation of the underlying distribution.

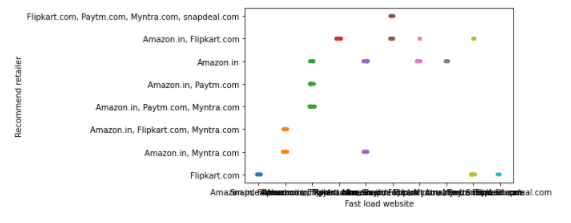


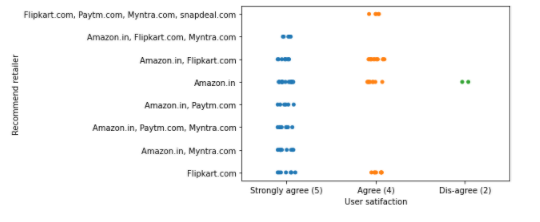












1. **Catplot**

This function provides access to several axes-level functions that show the relationship between a numerical and one or more categorical variables using one of several visual representations.



* **Interpretation of the Results**

We have some observations based on visualizations, pre-processing and modelling.

We will remove the outliers and make the data clean.

**Histogram** is used to check the frequency distribution of the data. We used histograms for *'* *Which of the Indian online retailer would you recommend to a friend', '* *Website is as efficient as before', '* *Change in website/Application design', '* *Longer time to get logged in (promotion, sales period)', '* *41 Monetary savings', '* *Presence of online assistance through multi-channel', etc.*

We have used **barplots, pointplots, jointplots** also.

Pairplot show the relationship between each variable with every other variable.

**CONCLUSION**

* **Key Findings and Conclusions of the Study**

So, our Aim is achieved as we have successfully ticked all our parameters as mentioned in our Aim Column. It is seen that the all the four models, KNeighboursClassifier, SVC, DecisionTreeClassifier and RandomForestClassifier performs well.

There is no definitive guide of which algorithms to use given any situation. What may work on some data sets may not necessarily work on others. Therefore, always evaluate methods using cross validation to get a reliable estimates.

In general, how we evaluate the model, how we plan to calculate the target, is all depend on the dataset and how we thought about it. We can predict the target by forming other parameters as groups, some categories, or create some new columns based on some logic.

The research results also provide tools for both real and virtual markets. On the other hand, application of the same model by targeting customers of the ready-to-wear sector can be an interesting field of study which can be used to reveal what customer do prioritize and/or value. A larger sample of countries from different nations would be appropriate to provide geographical or country level benchmark on the basis of organized ready-to-wear retailers to figure out potential differences in cultural perspectives. Carrying out similar research at other sectors and bringing out sector by sector comparisons can be used as an extensive area of research. In conclusion, this research has drawn attention to the progressive importance of customer retention sustaining efforts in ready-to-wear sector that is under intense competitive pressure.

.

* **Learning Outcomes of the Study in respect of Data Science**

There is certain outcome while creating the model. Visualization helps in estimating the relationship between the variables. Through visualization we got to know how data varies.

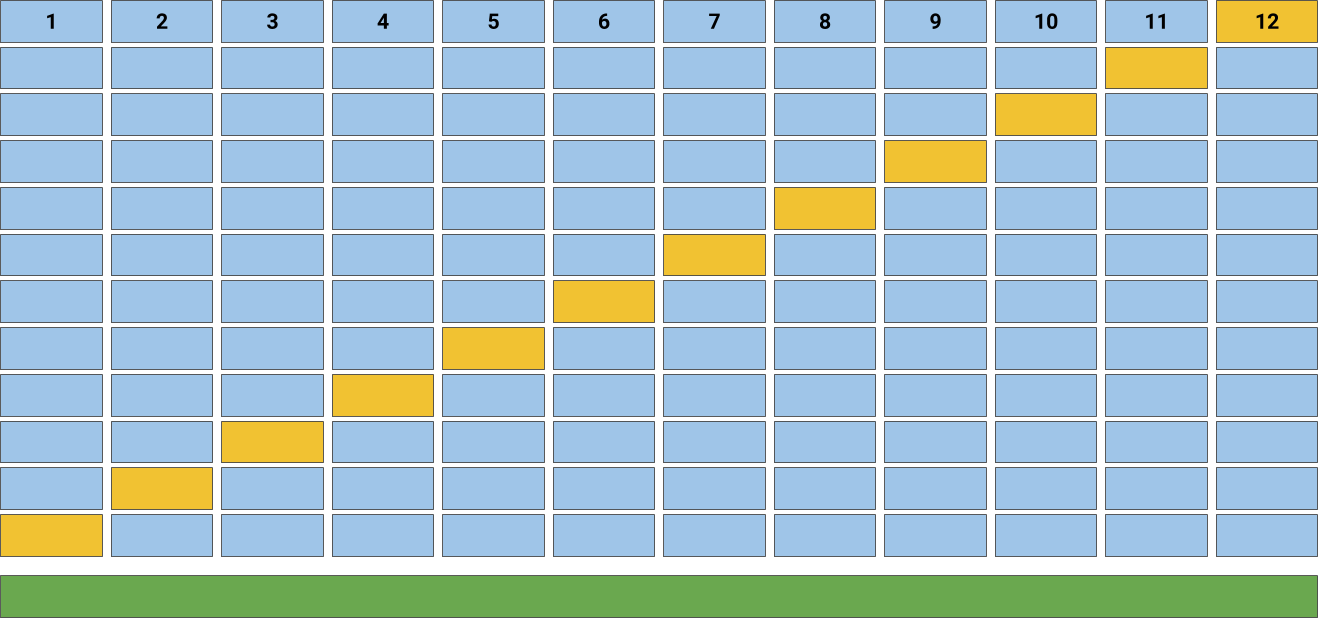
Boxplot shows us the outliers if present. Distribution plot shows the skewed data in the dataset if any. Histogram shows the frequency distribution of the data. Linear plot shows the linear relationship between the two quantities. Barplot shows the relationship in the form of bar. Pairplot shows the relationship between the entire variable at once only.

We have used various regression algorithms as mentioned below.

* KNeighborsClassifier
* GridSearchCV
* SVC
* DecisionTreeClassifier
* RandomForestClassifier

Generally we determine whether a given model is optimal by looking at its F1, precision, recall, and accuracy (for classification), or it’s coefficient of determination (R2) and error (for regression). However, real world data is often distributed somewhat unevenly, meaning that the fitted model is likely to perform better on some sections of the data than on others. Yellowbrick’s CVScores visualizer enables us to visually explore these variations in performance using different cross validation strategies.

Cross-validation starts by shuffling the data (to prevent any unintentional ordering errors) and splitting it into *k* folds. Then *k* models are fit on k−1kk−1k of the data (called the training split) and evaluated on 1k1k of the data (called the test split). The results from each evaluation are averaged together for a final score, and then the final model is fit on the entire dataset for operationalization.



* **Limitations of this work and Scope for Future Work**

The thesis is conducted as a case study by basing its findings and results on a specific area where the case company operates. For the competitive advantage reason, parts that include confidential information are hidden from another version of the thesis. Empirical data received from the interview and the focus group is not suitable for generalized usage due to its focused, one company orientation. A customer retention specific angle has been chosen for this thesis and it was decided to look at the concept from a certain company’s point of view. Due to the specific company focus with the research aim, the results and the final discussion around the topic have classified as confidential.