

CUSTOMER RETAINTION

*Submitted by:*

**SAYAN KUMAR BHUNIA**

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped me and guided me in completion of the project.

* [https://towardsdatascience.com/](https://towardsdatascience.com/job-salary-prediction-with-nlp-machine-learning-and-deep-learning-b87a96336b08)
* <https://anshikaaxena.medium.com/>
* <https://medium.com/https://medium.com/>

**INTRODUCTION**

* Business Problem Framing

Customer satisfaction has emerged as one of 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, theories 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 an 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.

**Analytical Problem Framing**

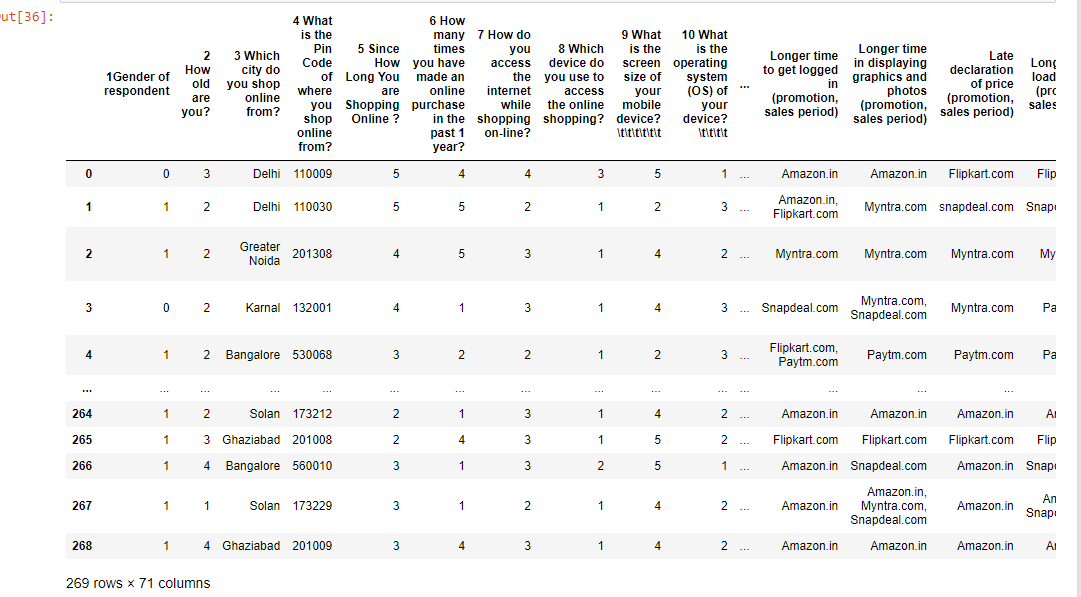
* Mathematical/ Analytical Modelling of the Problem

We shall build a supervised classification model to predict the risk of loan default.

Now, when we talk about building a supervised classifier catering to certain use-case, for example, classifying risk of loan default, following three things come into our minds:

* **Data** appropriate to the business requirement or use-case we are trying to solve
* **Optimize**the chosen model to ensure best performance.
* Data Sources and their formats

I am using excel format file which is having 269 rows × 71 columns.



* Data Pre-processing

Following steps have been performed on the data.

* **checking missing values**-
  + If there is any missing value present in your data set then for a better and correct accuracy you have to impute it.
  + If missing data present in object type column, then you have to take most frequent value for your missing data.
  + If missing data present in int or float type column then use mean/median for missing value.

In the following case no missing value present:

* **Encoding categorical variables** -as we can see there are 3 object data type columns present so we will encode it into (int) format.
  + Apply **Label Encoding**, if number of categories in a categorical variable is equal to 2.
  + Apply **One-Hot Encoding**, if number of categories in a categorical variable is greater than 2.

In the following case Label Encoding is used.



* **Feature scaling-** Feature Scaling ensures that all features will get equal importance in supervised classifier models. Standard scaler was used to scale all features in the data.
* **Reducing dimension of the data-** Sklearn’s pca can be used to apply principal component analysis on the data. This helped in finding the vectors of maximal variance in the data.
* **Outliers detection-** In simple words, an outlier is an observation that diverges from an overall pattern on a sample.

In the following case box plot is used to detect outliers.

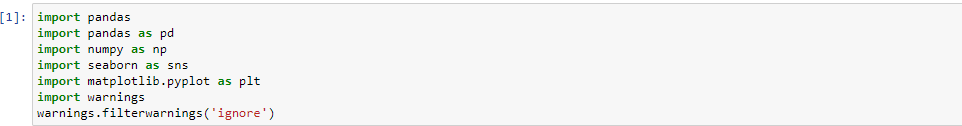
There are many types of outlier detection techniques such as Z-Score or Extreme Value Analysis, Probabilistic and Statistical Modelling, Information Theory Models, Standard Deviation etc.

* Outliers Removal

In our dataset, we observed variations in the relation between values of some attributes.

So that these types of rows are dropped from the dataset.

* Software Requirements and library Used





* **NumPy**

NumPy is a popular Python library for multi-dimensional array and matrix processing because it can be used to perform a great variety of mathematical operations. Its capability to handle linear algebra, Fourier transform, and more, makes NumPy ideal for machine learning and artificial intelligence (AI) projects, allowing users to manipulate the matrix to easily improve machine learning performance. NumPy is faster and easier to use than most other Python libraries.

* **Scikit-learn**

Scikit-learn is a very popular machine learning library that is built on NumPy and SciPy. It supports most of the classic supervised and unsupervised learning algorithms, and it can also be used for data mining, modelling, and analysis.

* **Seaborn**

Seaborn is another open-source Python library, one that is based on Matplotlib (which focuses on plotting and data visualization) but features Pandas’ data structures. Seaborn is often used in ML projects because it can generate plots of learning data. Of all the Python libraries, it produces the most aesthetically pleasing graphs and plots, making it an effective choice if you’ll also use it for marketing and data analysis.

* **Pandas**

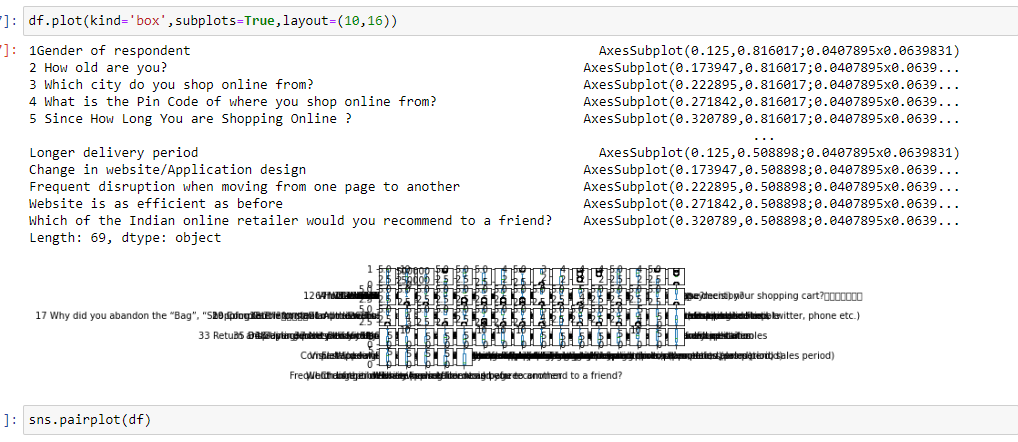
Pandas is another Python library that is built on top of NumPy, responsible for preparing high-level data sets for machine learning and training. It relies on two types of data structures, one-dimensional (series) and two-dimensional (Data Frame). This allows Pandas to be applicable in a variety of industries including finance, engineering, and statistics. Unlike the slow-moving animals themselves, the Pandas library is quick, compliant, and flexible.

* Class imbalance problem

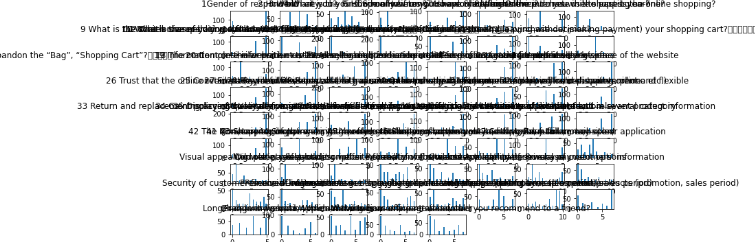
The first challenge we hit upon exploring the data, is class imbalanced problem. Imbalance data will lead to a bad accuracy of a model. To achieve better accuracy, we’ll balance the data by using Smote Over Sampling Method.

* Visualizations

For better understanding of outliers, I have used boxplot.



For better understanding of skewness, I have used distribution plot.



**CONCLUSION**

Key aspects of building successful classifier are:

* Selecting correct data according to the purpose or problem statement.
* Proper processing and understanding of the data
* Selecting the model and optimizing the model.

In this project I have dealt with outliers .

I have used label encoder for encoding object data type into int datatype as machine doesn’t understand object type data.