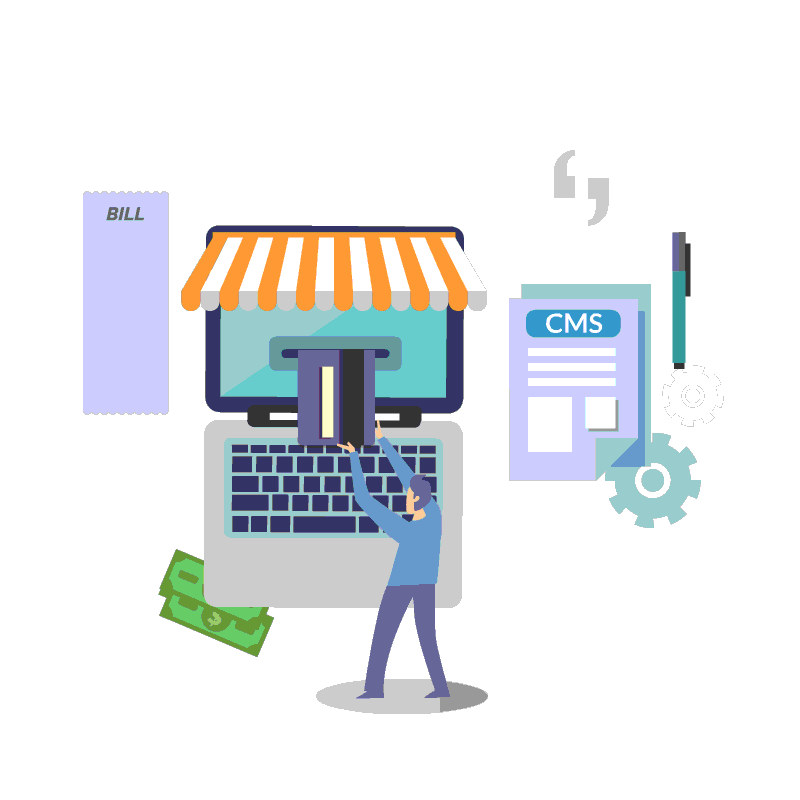
**Ecommerce Customer Churn Analysis and Prediction**

The technology has always been an instigating factor in progress for human civilization which resulted in driving the customer services to a greater need. The enrichment of technology has amplified and embellished the customer interaction among various business to consumer sectors. These technological upgrades have a huge impact on the retail industry which is an ever-growing market with key competitors around the world. In a consortium of multiple competitors in the same business, the re-engagement of disinterested customers is essential rather than winning a new customer. The sustenance of a customer can be figured out by Churn Prediction. Churn prediction is a new promising method in customer relationship management to analyze customer retention in subscription-based business. It is the activity of identifying customers with a high probability to discontinue the company based on analyzing their past data and behavior. It looks at what kind of customer data are typically used, does some analysis of the features chosen, and initiates a churn prediction model. Thus, churn prediction is a valuable approach in identifying and profiling the customers at risk.

# **OVERVIEW**

## **PROBLEM STATEMENT**

In the digital world, people want to purchase easily, quickly and within budget. In order to save time on shopping, they prefer online markets where they can find a huge variety of products at an affordable price that can be indemnified through the web and obtained at the doorstep. Marketing campaigns invite people to online shopping. In recent times the number of retailers online is increasing. Satisfying the customer is a tedious thing, as they want the best and reasonable product. In order to get a favorable product, there is a chance of switching to other online sites from classic, which leads to loss of customers. If this continues for a certain period of time it leads to customer churn.

Retailing gives an approach to products to find a good pace to customers. It endures snags like investment on labour, gratifying the bargaining customer. e-retail provides muddle free shopping to buyers, as they have various preferences.

To stay in a super competitive online market, retailers should undergo proceedings like verifying the identity of customers, being loyal and transparent to customers, following return and refund policies, and securing customer data. A customer likely to break the relationship or dwindle the purchase rate is known as churn. Customer churn occurs when a customer stops employing a retailer’s product, stops visiting a specific place of business, shifts to lower-tier experience or shifts to the contender’s products. Retailers need an abiding strategy to manage customer churn. Measuring the churn rate is kind of crucial for retail businesses because the metric reflects customer response towards the merchandise, service, 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 high 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.

One of the key methods to predict customer churn is machine learning. Specialized algorithms, used by companies, are adapted to specific problems and can perform such tasks as identifying obvious or latent features of customer’s behavior. It helps to understand better what are the reasons that people keep buying and what makes them leave. These algorithms can identify which buyers can become VIP and bring you most of the profit. Customer attrition can be done using different classification and predictive models. Efficient algorithms based on its accuracy are subjected to soft voting, that elect ensemble models as best to follow in impending works.

## **PROJECT OVERVIEW**

Customer churn prediction is highly important in e-commerce. The information about potential churners is valuable, because it allows you to take action when you still have time, and stop your customers from leaving. It is an essential action not only because you can lose additional revenue with every churn customer, but also because the buyer’s spending before the churn date might not cover the initial spending on acquiring this buyer. It’s a matter of fact, that retaining an existing buyer is cheaper than getting a new one. It costs five times as much to attract a new customer than to keep an existing one.

### OBJECTIVE

In this project, you aim to forecast the customer churn in ecommerce using various Machine Learning Algorithms to analyze if customers churn are impacted by different factors.

### MISSION

* Exploratory analysis of data to extract new insights into the consumer behaviours data-driven strategies through visual representation of the analysed data.
* Build a customer churn model based off of data to identify high risk customers and inform retention strategies and marketing experiments
* Analyse recency-frequency-monetary (RFM) analysis to model churn and customer lifetime value (CLV)

## EXPECTED OUTCOME

## 1. Solid consolidated project documents and scripts of your ML models engineering.

## 2. Be able to understand and practice data science workflow.

## 3. Confidence in using data science toolbox and Jupyter notebook report

## 4. Be able to pinpoint abnormalities in the data.

## 5. Able to produce a well-organized report which contains relevant insights.

6. Master the machine learning skill in building model prediction for this business case and focus heavily on how to apply and evaluate the data science methods, and to suggest business strategies in a real-world business setting after you analyze the given data.

## 7. A summary of your process" - as in thought and work process, the logic / critical thinking behind it, how you solved the problem, tools used and reasoning.

SUBMISSION OF CAPSTONE PROJECT

You will need to deliver three components:

1. A clean version of your code (to Github) with a README giving instructions about how to run it.
2. A document describing your technical solution and evaluation results. You need to submit it in PDF format to Github along with the code, in the root directory.
3. An oral presentation describing your technical solution and evaluation results. You are required to use slide presentation (i.e. PowerPoint, Google Slide, Notebook...etc) and also have to submit your presentation to Github.



For the code and document, which must be submitted to Github, here are the instructions:

1. Put the files for this assignment in a dedicated folder in the group repository you created.
2. Tag your commit with “final version of ML track”.
3. Double-check that Lymeng is added as a collaborator for this repository.
4. Send a link to the tagged commit by email to Lymeng (**lymeng@mekongbigdata.com**) with Sokhna (**sokhna@mekongbigdata.com**) in cc.

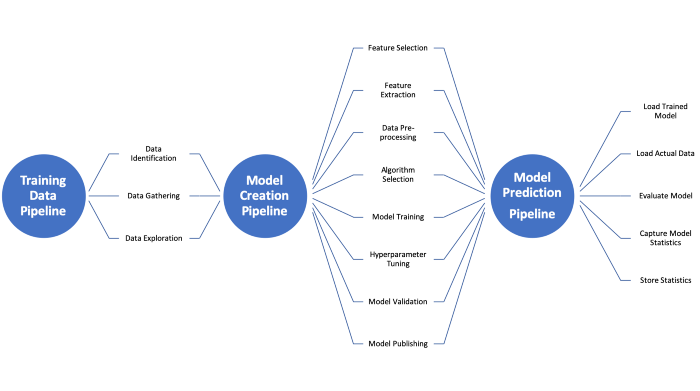
## DATA

Data Variable Description

* **CustomerID:** Unique customer ID
* **Churn:** Churn Flag
* **Tenure:** Tenure of customer in organization
* **PreferredLoginDevice:** Preferred login device of customer
* **CityTier:** City tier
* **WarehouseToHome:** Distance in between warehouse to home of customer
* **PreferredPaymentMode:** Preferred payment method of customer
* **Gender:** Gender of customer
* **HourSpendOnApp:** Number of hours spend on mobile application or website
* **NumberOfDeviceRegistered:** Total number of deceives is registered on particular customer
* **PreferedOrderCat:** Preferred order category of customer in last month
* **SatisfactionScore:** Satisfactory score of customer on service
* **MaritalStatus:** Marital status of customer
* **NumberOfAddress:** Total number of added added on particular customer
* **Complain:** Any complaint has been raised in last month
* **OrderAmountHikeFromlastYear:** Percentage increases in order from last year
* **CouponUsed:** Total number of coupon has been used in last month
* **OrderCount:** Total number of orders has been places in last month
* **DaySinceLastOrder:** Day Since last order by customer
* **CashbackAmount:** Average cashback in last month

## **MILESTONES**

This part serves as a suggested project planning. Some parts can be later adjusted through discussions with mentor



**Step 1 : Project Overview and Setup**

In this step you will get an overview of the project that you have to complete by the end of the program. You are going to set up and upload your data into your own work environment on a local machine.

**Step 2 : Study Data and Data Preparation**

The data provided in this project is relatively formatted. However, you may need to deal with some null values or transform some categorical variables. You are also encouraged to seek external data sources related to Ecommerce data. In this step you will focus on investigating and understanding the data by using visualization tools or libraries packages in Python. You also need to prepare and clean your data for the next module.

**Deliverables:**

* Look at the data variables and try to understand the meaning behind it.
* Investigate and clean the data.
* Study each variable in comparison with target variables.

**Step 3 : Exploratory Data Analysis**

In this step, you will focus on exploring the data by using statistical analysis in combination with visualization coded by libraries packages in Python. Based on the preprocessed data, you are expected to conduct a series of exploratory data analysis to get a taste of how the data looks or how it is distributed using histogram or other visualization or statistical methods to testify your data normality and satisfy machine learning assumptions. In this part you are required to perform various data visualisations to show what you have found and you also have to visualize your findings and to identify anomalies with python.

**Deliverable:**

* Create correlation between all relevant variables
* Use several different exploratory analyses to identify the key variables for your regression equation such as correlation plots, heatmaps, histograms etc
* Provide a statistical detail of data.
* Create visualization from the data with python.
* Arrange a coherent Jupyter-Notebook report.

**Step 4 : Feature engineering**

It is your job to perform feature selection, feature engineering, and discover the features that are indicative of someone paying or defaulting on their loan.

**Deliverables:**

* Use feature engineering where necessary.
* Use dimensional reduction where necessary.
* Clean and process data for machine learning.
* identify the right variables on which the analysis could be conducted

**Step 5 : Predictive Model for Classification**

Here comes the most exciting part! In this step, you will focus on making a predictive model based on supervised learning. Based on the previous data exploration and feature engineering, you can predict whether the transaction is analysed to be fraud based on some classification algorithms or various classification models so as to avoid the loss that is predicted to be a fraud transaction.

**Deliverables:**

* In this project, if you conduct multiple linear regression to predict the future sales. There are several different factors that you analyze in your regression model starting with a full model with all the variables and then moving towards a reduced model by eliminating insignificant variables.
* In Machine Learning, problems like fraud detection are usually framed as classification problems. Determine at least **3** ML algorithms that would have enabled you to have a better accuracy
* Arrange a coherent presentation report of the results, a well-organized Jupyter-Notebook report

**Step 6 : Model Comparison and Evaluation and Final Report**

Your goal is to compare and identify the best model for predicting fraud detection and justify why. Here you are expected to come up with a set of evaluation criteria to compare different model’s performance and discuss their confusion metrics and evaluate your model with different metrics and improve your model. It is important to try to explain why the models give such results and what are their pros & cons under certain constraints.

**Deliverables:**

* Compare ML algorithms by using accuracy metrics (precision, recall …)
* Your model should achieve more than 80% accuracy score
* Arrange a coherent presentation report of the results, a well-organized Jupyter-Notebook report

**Step 7 : Segmentation and Final Report**

In this step, you will focus on grouping the dataset into different segments based on preferences and other feature engineering. You also need to provide a recommendation for best actions that can be taken to improve revenue from various segments, cost reduction initiatives for the segments found, marketing campaign or new product opportunities..etc for each segment based on your findings.

For this step, please write a final report in the Jupyter notebook to include all your codes, graphs and writeup.

**Deliverable:**

* Using the ML algorithms to make predictions based on objectives of the project
* Give a meaningful representation for each cluster if there is required
* Provide recommendations