

**Capstone Project:** Customer Segmentation

# Machine Learning Track

## Overview

The purpose of this project is to provide the opportunity for students to explore and to have hands-on experience in applying ML learning models in practice. To become an expert machine learning engineer you need practice and experience. By completing this capstone project you will get an opportunity to apply the knowledge and skills in data analysis and ML modelling that you have gained throughout the series. This capstone project will test your skills in data cleaning, data exploratory, feature engineering, modeling, data wrangling, data organization, and machine learning.] The capstone is organized into 3 modules. Each module will take about 1-2 weeks (4-6 hours).

## Expected Outcome

1. A solid consolidated project documents and scripts.
2. Be able to understand and practice data science workflow.
3. Confidence in using data science toolboxes.
4. Be able to pinpoint abnormality in the data and extract useful information.
5. Define at least three (3) ML pipelines
6. Be able to transform, train models, and evaluate models.
7. Able to demonstrate data work and application to employer
8. Able to share the project output and application with public via blog, github, web dashboard or live predictions...etc

## Problem Statement

With the current competitive marketplace, it is essential to understand customer behaviour, their types, and their interests. Especially in targeted marketing, categorizing and understanding customers is a crucial step in forming effective marketing strategies. By creating customer segments, marketers can focus on one segment at a time and tailor their marketing strategies.

Customer Segmentation is one the most important applications of unsupervised learning. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base. Customer Segmentation is the process of division of the customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately. Companies aim to gain a deeper approach to the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer.

## Module 1

In this exercise, we will look into detail of the database containing online retail data of a UK-based registered online store. As an analyst, you are asked to produce a statistical baseline and to identify the abnormality with python language. In addition, you will need to prepare a Jupyter report to show the value that you have extracted from the data. Work together individually, explore the data with statistical methods by completing tasks below.

* Look at the data variables and try to understand the meaning behind it.
* Investigate and clean the data.
* Study each variable in comparing the purchase amount and customer.
* Create correlation between all relevant variables.
* Provide a statistical detail of data.
* Create visualization from the data with python.
* Arrange a coherent Jupyter-Notebook report.

**Duration:** 5 Hours (2 Weeks)

**Outcome:** A well-organized Jupyter-Notebook report

## Module 2

As a ML specialist, you are asked to perform data manipulation and data preparation for feeding the machine learning model in the next module. In addition, you will need to complete tasks below:

* Perform data wrangling and transformation.
* Clean and process data for machine learning.
* Prepare feature engineering.
* Determine and automate pipeline of feature engineering processes.
* Arrange a coherent presentation report of the results.
* Put the best model into production (optional)

**Duration:** 5 Hours (2 Weeks)

**Outcome:**  Aggregated data for clustering and a pipeline for feature engineering

## Module 3

In this task, you are asked to segment the customer data created above by using a machine learning algorithm (Hint: unsupervised learning). In addition, you will need to prepare a final report and presentation to showcase your capstone project with the completion of the task below.

* Use dimensional reduction where necessary.
* Determine the ML algorithms and optimal number of clusters to apply.
* Using the ML algorithms to find loan segmentation.
* Give a meaningful representation for each cluster.
* Provide recommendations
* Arrange and present a clear report.

**Duration:** 5 Hours (2 Weeks)

**Outcome:** Segmentation models, and a final report with presentation.

## **Submitting** 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 PPTX 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 ([**@**](https://github.com/tholop)**LymengLM**) 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

| **Desc** | **Value** |
| --- | --- |
| *Data Name* | Online Retail Data Set of UK-based Online Store |
| *Total Attributes* | 8 |
| *Total Instances* | 541,909 |
| *Record Date* | 06 November 2015 |