**Project Proposal**

**Project**: Customer Segmentation using Unsupervised Learning

**Team Lead**: Ramkumar M

**Team Members**: Apoorva Bhatla, Tushar Joshi, Sai Nitheesh Kumar Reddy, Ramanpreet Kaur, and Zahoor Ansari

# Timeline: 14 working days

**Planned Start Date:** 9th July 2021

**Planned End Date:** 29th July 2021

**Recommended Preparation and Study Material**: Customer Segmentation, Clustering and its algorithms: K-means, DBSCAN, Visualization libraries, EDA

Following publication from Packt can be useful for Students in this project

* [The Unsupervised Learning Workshop](https://courses.packtpub.com/courses/unsupervised-learning)
* [Applied Unsupervised Learning with Python](•%09https:/www.packtpub.com/product/applied-unsupervised-learning-withpython/9781789952292)
* [Hands-On Unsupervised Learning with Python](https://www.packtpub.com/product/hands-on-unsupervised-learning-withpython/9781789348279)
* [Python Data Science Essentials - Third Edition](https://www.packtpub.com/product/python-data-science-essentials-thirdedition/9781789537864)

**GitHub Repo Link**: <https://github.com/TeamEpicProjects/Credit_Card_Customer_Segmentation>

**Dataset Link**: [data\_credit\_card\_customer\_seg.csv](https://github.com/TeamEpicProjects/Credit_Card_Customer_Segmentation/blob/main/data_credit_card_customer_seg.csv)

**Recommended System Setup:**

# Hardware Requirements

* System with minimum 8 GB RAM is recommended

# Software Requirements

* Python Version 3 and above (Anaconda Distribution) - <https://www.anaconda.com/products/individual>
* Jupyter IDE Notebook (Will be part of Anaconda Download)
* Machine Learning Libraries in Python – Scikit Learn (Will be part of Anaconda Download)
* If additional package required, pip install can be done

# Problem Statement

RBL’s marketing department collects various customer specific data of the credit card holders. They need a mechanism to segment the customer based on underlying characteristics and form market clusters which will be easy for them to target and provide product ideas to the management. Currently these tasks are performed manually by trusting the judgment of experts in the field. These can lead to human error, biased decision and other factors which may not be helpful to create a customer cluster that actually exist. They want to design a system that would automate this process and help the different stakeholders to make informed business decision.

# User Stories

* As a team lead, I would like to check the data quality and feasibility for implementation of clustering algorithm
* As marketing lead, I am interested to check if the data has captured necessary features and how they are distributed
* As business stakeholder, I am interested to check the overall progress of the project which includes data requirements
* As Marketing Manager, I am interested to see if there are any initial business insights from EDA
* As Data science team lead, I want to check and validate the iterations on various cluster algorithm to check and identify the suitable one for the problem statement
* As marketing head, I need to see the cluster formation and how it is related to real time data, visualization of cluster will be more helpful
* As business stakeholder, I am more interested to get actionable insights from the exercise and observations that impacts business
* As a bank manager, I would like to see the different clusters of customers with the pattern of spending habits.

# Expected Solution

Solution focus on developing a clustering algorithm based on unsupervised learning. There are various clustering algorithm, and the task to identify the best one for the current scenario and build and fine tune them based on characteristic

**Milestones**

**Milestone 1**

* Data understanding and initial observations
* Data preparation which involves cleaning, encoding, imputation etc
* Exploratory Data Analysis
* Statistical approach to understand the data feature relations and properties
* Check to perform if dimensionality reduction is necessary
* Identification of ideal clustering algorithm for the problem statement and dataset

**Milestone 2**

* Application of optimization techniques to identify the best parameters for the cluster formation
* Implementation of clustering algorithm to segment the data
* Validating the cluster formation
* Aligning with real data to create cluster customer map
* Visualization of clusters
* Providing observation based on the outcome

# Work Package

**Note:** Overall project output will be based on group work and collaboration among team members. However, there are areas where each individual team member can enrich the analysis by extending the approach or performing more iterations in EDA, model building etc. This is highlighted below, and each individual team member will be measured on the contribution they provide

|  |  |  |  |
| --- | --- | --- | --- |
| Day | Work Package | Group Task | Individual Task |
| 1 | Initial Data understanding   * Get Overview of the Project from Mentor * Access the data and import it in Jupyter Notebook * Check the features and have a basic understanding of the data * Provide your initial observations in the Notebook | Completion of initial analysis of data   * Types of features * Initial observations and recommendation on Data preprocessing * Implementation in the Notebook | Each individual will come up with idea on how they want to Preprocess data   * List atleast three approaches on Data preprocessing * Atleast 3 observations on data provided * Collaborate with team on providing a consolidated output |
| 2 | Data understanding in detail   * Perform Exploratory Data Analysis on the dataset * Create bivariate multi variate plots and list down observations on feature relationship | Completion of EDA with analysis based on   * Distribution * Feature relation * Multi collinearity check * Other data properties based on statistical tests | * Divide the EDA steps among each member and perform and consolidate the work * At least 2 EDA approach implemented with proper annotations and observations |
| 3 | Data Preparation   * Check on the data quality and any issues in data * Perform data cleaning steps * Perform encoding and data modifications if any * Based on EDA / observation perform any corrective actions | Completion of data preparation steps   * Data cleaning and preprocessing * Encoding if any * Outlier treatment if required * Any corrective action based on EDA observations | * Team members can split the work and complete atleast one Data preparation step * Collaborate for consolidation of work |
| 4 | Feature Creation / Selection   * Check if Dimensionality reduction is required * Make two approach, one with PCA implemented and other with normal data * Check if any feature selection method is required for normal data | * Check for feature creation or selection possibilities * Perform iterations to get best possible feature combinations for the model | Perform atleast one iteration and provide observations  Collaborate to consolidate the findings |
| 5 | Initial Model Creation and feasilibty check   * Implement set of base clustering algorithm and check the output * Based on this provide your observation and the choice of clustering algorithm | * Identification of best model approach for the dataset * Perform iterations to find the best model approach and consolidate the observations | Perform atleast one iteration and provide observations |
| 6 | * Provide a short presentation of the implementation outcome   Explanation of approach next | Collaborate and provide a single presentation for the interim report | Contribute to the presentation work |
| 7 | * Selection of the clustering method and performing optimization steps * Finding the optimal cluster for the dataset | Group work | NA |
| 8 | * Performing cluster evaluation and reporting the cluster quality formation * Providing observation on the output | Perform cluster evaluation with various metrics and approach available | * Perform cluster evaluation with atleast one metric or approach and provide observation * Collaborate to consolidate the observations |
| 9 | * Cluster visualization using appropriate plotting techniques | Group Work  Perform cluster visualization to dissect the formation and to visualize the pattern formed using one or more approach | NA |
| 10 | * Aligning with real-time data to understand the cluster formation and providing business insights |  | * Contribute with ideas and business insights * Atleast 2 ideas need to be contributed |
| 11 | * Additional time to complete the work items | NA | NA |
| 12 | * Completion of final report and presentation | Group work | NA |
| 13 | * Additional time to complete the work items | NA | NA |
| 14 | * Final presentation of the result and outcome * Answer business relevant questions | Present the final outcome and answer the business relevant question | Actively participate to answer questions based on presentation and business from stakeholders |

**Assessment criteria for the solution**

**Note:** The individual work items will be used for evaluation of each team members and these form the metrics. Weightages are allocation for each work items and the outcome accordingly.

Solution is measured on the following factor and metrics:

* Exploratory Data Analysis with actionable insights
  + Quality of the EDA plot implementation and annotation
  + EDA covering facets of data including distribution, feature relation etc.
  + Statistical approach to perform test on data features
* Data Preparation
  + Data preparation should enable the model run without error
  + Generation of features / feature selection / elimination / Dimensionality reduction
* Identification of optimal Model by following best practices
  + Model should generate clusters from the data
  + Application of optimization of model for finding best cluster
* Measure of cluster quality
  + Use of cluster evaluation method such as silhouette method or other metrics to check the cluster formation quality
  + Metrics should yield a positive value above 0.4
* Quality of cluster visualization
  + Cluster visualization should show the cluster formation clearly in the plot
* Actionable insights based on cluster formation
  + Provide at least 3 actionable insights based on aligning with business context
  + Provide recommendation based on insights