Customer Segregation with Data Science

T A Muhammad Tawfeeq 211521243102

Phase 1

Problem Definition

The problem at hand is to effectively segregate customers using data science techniques in order to enhance the customer experience, tailor marketing efforts, and optimize business strategies based on their behaviour, preferences, and demographic attributes. Design thinking is a human-centred approach to innovation that focuses on understanding the needs of users and developing solutions that meet those needs. Data science is a field of study that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data. The goal is to create distinct customer segments based on relevant attributes and behaviours, allowing for personalized interactions and targeted campaigns.

Design Thinking

Data Collection and Integration:

Gather and integrate data from the previous data of customer purchase and their way of payment. Ensure data accuracy, consistency, and compliance with privacy regulations.

Data Pre-processing and Cleansing:

Cleanse and pre-process the collected data, handling missing values, outliers, and data quality issues. Normalize and standardize the data for consistent analysis. Modifying the data to be suitable for the model we about to use to segregate the customers.

Feature Engineering:

Identify relevant features and create new variables that may enhance the model's predictive capabilities. Incorporate external data sources and demographic information, if applicable.

Clustering Algorithms:

Develop a preliminary framework for customer segmentation based on the identified data points. Utilize data science tools and techniques such as clustering

algorithms (e.g., k-means, hierarchical clustering) to create initial customer segments. Validate the prototype with a subset of customer data to ensure its effectiveness.

Visualization:

Perform EDA to gain insights into historical company registration trends.Visualize data patterns, correlations, and anomalies. Identify key factors that influence company registrations, such as economic conditions, industry trends, or regulatory changes.

Model Validation and Performance Metrics:

Establish validation procedures to assess the accuracy and reliability of the predictive models. Define appropriate performance metrics, such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), or R-squared, for model evaluation.

Real-time Data Updates:

Implement mechanisms for real-time or periodic data updates to ensure that the predictive models stay current and relevant.

User Interface and Reporting:

Create a user-friendly interface for stakeholders to access and interact with the predictive system.
Generate reports and visualizations to communicate predictions and insights effectively.

Scalability and Efficiency:

Design the system to handle large volumes of data efficiently and accommodate scalability as data grows over time.

Feedback Loop:

Establishing a feedback loop with users to collect ongoing input and make iterative improvements to the system.

Interpretation and Analysis:

Observe patterns, clusters, or groupings in the reduced-dimensional space. Analyze the relationships between data points and identify any trends or anomalies. Optionally, color-code points based on known customer attributes (e.g., demographics) to gain further insights.

Phase 1 Deliverables:

For Phase 1, the following deliverables are expected:

Clean and Preprocessed Dataset:

Provide a dataset where duplicates have been removed, missing values have been addressed, and categorical variables have been encoded. This dataset should be ready for use in machine learning models.

List of Engineered Features:

Document all the newly created features along with their descriptions and how they were calculated. This list should make it clear how each engineered feature adds value to the predictive model.

Data Preprocessing and Feature Engineering Report:

Create a detailed report that outlines the steps taken during data preprocessing and feature engineering. This report should cover the methods used, challenges encountered, and the rationale behind the decisions made at each step.

In Phase 2:

Incorporating dimensionality reduction techniques like Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbour Embedding (t-SNE) can be highly beneficial for visualizing high-dimensional customer data and identifying underlying patterns.

Conclusion:

In Phase 1, we have established a clear understanding of our goal: to segregate customers with data science. We outlined a structured approach that includes data source selection, data preprocessing, feature selection, model selection, model training, and evaluation. This sets the stage for our project's successful execution in subsequent phases.