ABCDEats Inc.

Data Mining Project Guidelines

version: v8.2024.09.16

Fall Semester 2024-2025

Master in Data Science and Advanced Analytics

NOVA Information Management School

Introduction

Consumers today are becoming more selective about the businesses they support and where they spend their money. This makes it essential for companies to develop a deep understanding of their customer base in order to tailor their services and marketing strategies more effectively.

Customer segmentation is the process of dividing customers into smaller groups that share similar characteristics or behaviours. In the case of a food delivery company for instance, some groups may prioritise convenience and quick delivery, while others might focus on price or healthy options. Recognizing these differences will allow the company to better customise their products, services, and marketing efforts to meet the specific preferences and needs of each segment, enhancing customer satisfaction and ultimately boosting retention and revenue.

Project Description

In this project, you will act as consultants for ABCDEats Inc. (ABCDE), a fictional food delivery service partnering with a range of restaurants to offer diverse meal options. Your task is to analyse customer data collected over three months from three cities to help ABCDE develop a data-driven strategy tailored to various customer segments. The description of the data is provided under the Dataset Description section of this document.

We recommend segmenting customers using multiple perspectives. Examples of segmentation perspectives include value-based segmentation, which groups customers by their economic value; preference or behaviour-based segmentation which focuses on purchasing habits; and demographic segmentation which categorises customers by attributes like age, gender, and income to understand different interaction patterns.

Ultimately, the company seeks a final segmentation that integrates these perspectives to enable them to develop a comprehensive marketing strategy.

Key Dates

- Project Part 1 Due: 04 November 2024 21h00
- Project Part 2 Due: 03 January 2025 21h00
- Project Discussion: 21-25 January 2025 (**not final**)

Project Evaluation

The project will be evaluated based on three components. All components are mandatory.

Part 1: Exploratory Data Analysis (20%)

Expected Outcomes

- Conduct an in-depth exploration of the dataset. Summarise key statistics for the data, and discuss their possible implications.
- Identify any trends, patterns, or anomalies within the dataset. Explore relationships between features.
- Create new features that may help enhance your analysis.
- Use visualisations to effectively communicate your findings.

Submission deadline

- 04 November 2024 21h00
- 10% penalty for each day of delay

Deliverables

- Written report, maximum of 5 pages.
- Python code, in a Jupyter notebook with cells already executed.
- See Project Deliverables section for details

Part 2: Final Report (70%)

Expected Outcomes

- Preprocess the data. Invest time into evaluating your preprocessing pipeline, explaining the choices you made and the advantages and disadvantages of different decisions.
- Justify the clustering approach. Determine and provide a rationale for the clustering solution, including the number of clusters, that you decide to use.
- Explain the clusters in your final segmentation. Analyse and describe the characteristics of each group, taking into account the perspectives you used. Create profiles that highlight the distinguishing features of each cluster.
- Suggest business applications. Based on your insights, define general marketing approaches for each cluster.

Submission deadline

- 03 January 2025 21h00
- 10% penalty for each day of delay

Deliverables

- Written report, maximum of 10 pages.
- Python code, in a Jupyter notebook with cells already executed.
- See Project Deliverables section for details

Part 3: Project Discussion (10%)

After submitting the final reports, each group will be scheduled for a 15-20 minute discussion with one of the instructors. During this discussion, students will answer questions about their analysis and explain their methodologies.

Each student will receive an individual grade based on their contribution to the discussion. This aims to assess each student's level of engagement with, and comprehension of the delivered work.

Target schedule: 21-25 January 2025. **This is not yet final** and will be confirmed later on Moodle.

Optional: Opportunity for Bonus Points (max 20%)

Note: *This does not replace the requirement to deliver the other parts of the project.*

As an optional part of the project, you may develop an interactive application that allows ABCDE to explore and interact with the EDA and customer segmentation analysis performed previously. The application should be user-friendly, visually intuitive, and allow the user to gain insights from the customer segmentation results.

This is intentionally open-ended; some suggestions are provided below, but the kinds of visualisations or features to include is limited only by your imagination and skill.

The only restriction is that it must be developed using Python.

Some recommended Python libraries:

- Bokeh
- Plotly

Possible application features:

- **Cluster Exploration**: The application must provide an interactive interface where users can explore the different customer segments. Each cluster should be clearly defined, with key demographic, behavioral, and preference-based characteristics displayed.
- **Visualisation Tools**: Implement visual tools such as charts, graphs, or scatter plots that help users visualize the clustering results. Consider using heatmaps, pie charts, or bar graphs to represent segment data.
- **Filter Functionality**: Allow users to filter clusters based on various attributes (e.g., age, location, purchase history) and explore how these filters affect the composition of the clusters.
- **Cluster Comparison**: Enable users to compare different clusters side-by-side in terms of their key characteristics, providing detailed insights into their similarities and differences.
- **Personalized Insights**: Include an option where users can input specific customer attributes (e.g., age, region, payment method) and get insights into which cluster the customer is most likely to belong to.

Submission deadline

- 07 January 2025 21h00
- **Please inform me by email if you intend to deliver this optional part.**

Deliverables

- Summary describing main features of the application, maximum of 3 pages.
- Python code used to develop the application.

Project Deliverables

Project report

- The report must be written in English.
- The report must be delivered in PDF format using the template provided.
- The report must contain the names and student numbers of ALL members of the group on the cover page.
- The report for Part 1 must not exceed 5 pages of content (excluding the cover page, index, annexes, appendices).
- The report for Part 2 must not exceed 10 pages of content (excluding the cover page, index, annexes, appendices).

Python code

- The submission of all Python code developed is mandatory.
- The notebook must include a Markdown cell (text cell) containing the names and student numbers of ALL members of the group.
- We are not evaluating the complexity of the code. Instead, the notebook should be easy to follow with the workflow structured logically, and include appropriate comments explaining key steps and decisions.

File naming guidelines

Part #	Report	Notebook
Part 1	`DM2425_Part1_99.pdf`	`DM2425_Part1_99.ipynb`

Part 2	`DM2425_Part2_99.pdf`	`DM2425_Part2_99.ipynb`
		
		In the case of multiple notebooks, name them in order and submit as a zip file; e.g.:
		
		`DM2425_Part2_99.zip`
		`- DM2425_Part2_99_01.ipynb`
		`- DM2425_Part2_99_02.ipynb`
		`- DM2425_Part2_99_03.ipynb`

*99 is the group number

Deviations to the guidelines may result in a deduction.

General Policies

Group composition

- Maximum of FOUR (4) students in each group. We recommend a group size of three (3).
- ALL students must be enrolled in a group on Moodle, regardless of group size.
- Students must be enrolled in a group on Moodle before the first delivery deadline.
- Changes to the group composition after the first delivery date is allowed only in the case of an existing group requesting to be divided.

Plagiarism check

All submitted reports will undergo a plagiarism check. Ensure that your work is original and properly cite any external sources used.

Use of Generative AI tools

- The use of generative AI tools, such as ChatGPT, is permitted but must be fully disclosed. It is essential that the students' own contributions to the report exceed that of any AI tools used.
- **Do not just copy and paste the results if AI tools are used.**
- Students must include a section in the annex of the report disclosing the use of AI tools. For example: *"ChatGPT was used to ... in sections X, Y, Z of this report*"; *"ChatGPT was used to generate the code to do ..."*
- Students may be asked to explain the meaning of any ChatGPT-generated content submitted to ensure comprehension.

Students are fully responsible for the contents of the report they submit, including any material generated or assisted by AI tools.

References

Dolnicar, S., Grün, B., & Leisch, F. (2018). *Market Segmentation Analysis*. https://doi.org/10.1007/978-981-10-8818-6

Jolaoso, C. (2024). Customer segmentation: the ultimate guide. https://www.forbes.com/advisor/business/customer-segmentation/

Dataset Description

The dataset provided for this project includes customer data from ABCDEats Inc. and consists of multiple columns capturing various customer attributes and aggregating their behaviour over a three-month period.

Each row corresponds to one customer, and the column descriptions are given below:

#	Column Name	Description
1	customer_id	Unique identifier for each customer.
2	customer_region	Geographic region where the customer is located.
3	customer_age	Age of the customer.
4	vendor_count	Number of unique vendors the customer has ordered from.
5	product_count	Total number of products the customer has ordered.
6	is_chain	Indicates whether the customer's order was from a chain restaurant.
7	first_order	Number of days from the start of the dataset when the customer first placed an order.
8	last_order	Number of days from the start of the dataset when the customer most recently placed an order.
9	last_promo	The category of the promotion or discount most recently used by the customer.
10	payment_method	Method most recently used by the customer to pay for their orders.
11	CUI_American, CUI_Asian, CUI_Chinese, CUI_Italian, etc.	The amount in monetary units spent by the customer from the indicated type of cuisine.
12	DOW_0 to DOW_6	Number of orders placed on each day of the week (0 = Sunday, 6 = Saturday).
13	HR_0 to HR_23	Number of orders placed during each hour of the day (0 = midnight, 23 = 11 PM).