MARCH, 2024

BUDGETING FOR THE FUTURE: RESIDENT LIFESTYLE EXPLORATION TO EMPOWER CITY COUNCILS

GROUP PROJECT
DATA MINING II 2023/2024

01

I. INTRODUCTION

Participatory budgeting stands as a key strategy in **municipal governance**, addressing challenges of citizen engagement and resource allocation. Since it is a democratic approach to decision-making, it empowers community members to actively shape the distribution of public funds, fostering transparency and accountability within local government structures. By involving residents in the budgeting process, participatory budgets not only ensure that public resources are allocated in line with community needs and priorities but also cultivate a sense of ownership and trust in governmental institutions.

Therefore, to reflect the needs and aspirations of all citizens, a certain city council should have a comprehensive understanding of its population before initiating a proposal for participatory budgeting.

II. PROJECT GOALS

The **city council** of **Mining City** is **concerned** about the **effectiveness** of its **participatory budgeting** process. Therefore, the council members have decided to collect extensive data and surveys, to gain valuable **insights** into the **demographics** and **lifestyles** of its **residents**. As a result, were identified **key labels** such as **'Travel Enthusiast', 'Health-Conscious', 'Adventure Seeker', 'Fitness Enthusiast',** and **'Investor'**.

The council intends to utilize this helpful information to make better decisions regarding how to allocate funds. They aim to **incorporate predictive modeling techniques** into their **budgeting process** and have sought **assistance from NOVA IMS**. They require **you and your team to address specific points:**

1. **Identification of Relevant Variables**: Selecting variables from the dataset that are pertinent to the participatory budgeting scenario.

2. **Multiclass Classification:** Using predictive modeling algorithms to forecast the preferences and priorities of various segments within the community ***(lifestyle_type)*** to provide city council members valuable tools to enhance the efficient and appropriate allocation of resources according to the population's needs.
   This will be completed through a Kaggle competition and the score of predictions will be evaluated using **F1 Score 'weighted'**.

02

III. DATASET

You have access to two different datasets:

In the **training set**, you will find the features that give information related to each citizen record. Use the training data and its features to build and validate your machine-learning models. The goal will be to use the models you created and make predictions on unseen data (i.e. your test set). **Important note: You should not consider the target variable as feature for any of the predictive models.**

In the **test set**, you will still have access to the same descriptive attributes associated with each citizen. However, you will not have access to the target variable of the multiclass problem.

The available data contains the following attributes:

ATTRIBUTE	DESCRIPTION
citizen_id	Unique identifier of the citizen.
Name	First name of each citizen.
Title	Title of each citizen.
date_of_birth	Date of birth of each citizen.
city	Name of citizen´s city.
country	Name of citizen´s country.
last_year_avg_monthly_charity_donations	The average of monthly charitable donations made by each citizen in the last year.
environmental_awareness_rating	A rating [0, 10] of each individual's awareness of and engagement with environmental issues.
financial_wellness_index	An index indicating each citizen´s overall financial health.
investment_portfolio_value	The value, in thousands of units of currency, of each citizen´s investment portfolio.
investments_risk_appetite	A measure of each individual's willingness to take risks in their investments.
investments_risk_tolerance	A measure of each individual's tolerance for risk in their investment choices.

03

ATTRIBUTE	DESCRIPTION
tech_savviness_score	A score representing each citizen´s proficiency and comfort with technology.
social_media_influence_score	A score representing each citizen´s influence and activity on social media platforms.
entertainment_engagement_factor	A score representing each citizen´s engagement with entertainment activities.
avg_monthly_entertainment_expenses	The monthly expenditure on entertainment for each citizen, in units of currency.
avg_weekly_exercise_hours	The average number of hours each citizen spends on exercise weekly.
health_consciousness_rating	A rating [0, 10] of each citizen´s awareness and proactive behavior towards their health.
stress_management_score	A score indicating how effectively each citizen manages stress.
overall_well_being	A score indicating each citizen overall status.
lifestyle_type	A categorization of the predominant lifestyle choice for each citizen **(Target Variable)**.

04

IV. OUTLINE

Your project deliverables (especially the report) should respect the following outline:

Abstract
A small summary of your work (200 to 300 words). The abstract should give an overview of your work: What is the context? What is your main hypothesis? What did you do? What were your main results, and what conclusions did you draw from them?

I. Introduction
- Overview of the project
- Main goals of the project (being a requirement for the course does not count)
- Did you find any research with similar objectives? What has been done? What did other researchers find? What would you expect your results to be based on their previous findings?

II. Data Exploration and Preprocessing
- Description of data received
- Steps taken to clean and prepare the data

III. Multiclass Classification
- Additional preprocessing steps adopted
- Description of the actions taken
- Results and discussion of main findings

IV. Conclusion
- Summary of the findings
- Do the findings match what you initially expected? How?
- Discussion of limitations of your work (e.g. what could you have done differently)
- Suggestions for possible work to follow on your work.

Note:
  - The body of text should only include Figures and Tables that are essential to understand your work, and therefore referenced in the text. Supporting figures and Tables can be added to Annexes.

V. DELIVERABLES

Upon the project's deadline, you will be required to submit:

- A Jupyter notebook (or a zip of multiple notebooks) featuring all the code you used, throughout the project, regarding your best scored solution.
  - The file naming format should follow ***XT_DM2_GroupXX_Notebook***, where ***XT*** should be DT or NT, depending if you are from Daytime or Nighttime classes, ***GroupXX*** should be your group number.

- A report that describes the analytical processes and the conclusions obtained with, at most, 10 pages (excluding cover, abstract and annexes).
  - The file naming format should follow ***XT_DM2_GroupXX_Report.pdf***, where ***XT*** should be DT or NT, depending if you are from Daytime or Nighttime classes, ***GroupXX*** should be your group number. The report should follow these settings**:**
    - **Heading 1: Calibri, Size 14 pt, in bold**
    - **Heading 2 (if needed): Calibri, Size 13 pt, in bold**
    - **Text: Calibri, Size 11 pt, line spacing of 1.15 pt and paragraph spacing of 6 pt**

VI. EVALUATION

Your work will be evaluated according to the following criteria:

CRITERIA	PERCENTAGE (%)	MAXIMUM GRADE (OUT OF 20)
Report Quality and Storytelling	20	4
Preprocessing and EDA	20	4
Multiclass Classification	20	4
Model performance on Kaggle	15	3
Creativity and Other Self-studies	10	2
Presentation	5	1
Discussion	10	2

06

Your grade will reflect our assessment of the quality of your work in terms of quality of writing, clarity, conciseness, correctness and efficiency. Please find below more details about what is taken into account for each topic:

- **Report Quality and Storytelling (4v):** A good report should, by itself, give the reader a clear picture of the problem you are tasked with, the steps you took, the rationale behind those steps, your main results and your insights. When referencing a figure, ensure you direct the reader's attention to the point you want to convey. This section also encompasses the overall quality of your introduction and conclusions.
- **Preprocessing and EDA (4v):** Describe the data and extract meaningful insights that you consider helpful. Avoid adding visualizations and elements that add nothing to address the problem at hand. This section also covers the initial preprocessing of the dataset. In essence, it should unambiguously explain what you did on a specific step and what are the reasons for your choice of approach.
- **Multiclass Classification (4v):** Describe your strategy for the multiclass classification objective. This section is separated into different components:
  - Additional Preprocessing (includes feature selection): 1v
  - Modelling approach (model assessment (holdout, cross-validation, etc...), algorithms used): 1v
  - Performance assessment (choice of metrics and interpretation of results): 2v
- **Model performance on Kaggle (3v):** Grading based on groups ranking order.
- **Creativity and other self-studies (2v):** This topic includes applying different techniques and aspects of creativity, such as choice of visualizations, approach or techniques used. If techniques not given during practical classes are used, you should provide a theoretical explanation for them in the annexes.
- **Presentation (1v):** In this section, we will evaluate the overall quality of your presentation as a standalone document.
- **Discussion (2v):** This section evaluates how you, as a group, deliver the message you want to convey whilst presenting it. How comfortable you are in answering questions about your data and methods is also under evaluation.

07

VII. PARTING NOTES

- **For modelling purposes, any algorithm implementation outside the vanilla scikit-learn is explicitly off-limits. Moreover, using Lazy Predict or similar AutoML packages is also not allowed.**
- The report will be the primary method of evaluating your work. When preparing it, remember that a reader should be able to understand your work without needing to check your notebook. We won't be able to consider any steps or results not mentioned in your report.
- Please don't provide long theoretical explanations of topics covered in class in your report.
- Everything featured in your report must have a clear purpose. Avoid including irrelevant/unimportant/redundant information, as the space is limited, and you will need it.
- Trustworthiness of the information you provide is key. You should look to source information you provide from peer-reviewed journals (thus, avoid citing Medium, TowardsDataScience and similar sources).
- Before submitting, run your notebook from the start one last time (if you used a GridSearch, you can comment this cell, but you should run the final model with the GS parameters in a different cell).
- All the unneeded code you used to obtain your final solution **should be part of your submitted notebook, but it should be commented.**
- We will run your Jupyter Notebooks if we have any doubts. So, please make sure we can run the notebook from start to finish in one go. Notebooks that do not fulfil this condition will be penalized.
- The report and code will pass through a process of plagiarism and AI generation checking.
- You must submit your predictions on the Kaggle competition to get points for that component.
- When determining the grade for your work, there will be a comparative component between your work and the works presented by your peers.

Friendly Reminders:

- Attendance at the defense is mandatory for approval in the project. The defense has a group component and an individual component.
- **If something is good enough to be mentioned in the report, it is also good enough to know. DO NOT include techniques/algorithms/steps you cannot explain in your report: we may (and probably will) ask about them in the defense.**
- Finished is better than perfect.