Group Project

Maximum points: 40 Project report due date: 29th Nov. 2024 at 23:00

Guidelines

- Group Project: This is a group project. Please communicate with your group members to complete the project on time.
- Problem Understanding: There are two parts to this project. Carefully read the problem description of both the parts before starting your analysis.
- Choice of Software: You are free to use either R or Python to perform your analysis.
- Code Requirements: Your code should be well-structured and free of errors. Please use clear and descriptive variable names.
- Deliverables: Specified separately for Part A and Part B (see below).
- Plagiarism Warning: Ensure that all the code and analysis are your original work. Plagiarism will lead to disqualification and academic consequences.
- Late Submissions: Late submissions will be accepted, but penalized.
- Points Distribution: The distribution of points and grading rubric for the questions is provided at the end of this document.
- Honor Code: Please include the following honor code (text in blue) on the cover page of the project report (Part B).

By submitting this assignment, we affirm the following:

1. If we have used AI tools like ChatGPT, Co-Pilot, etc., we only sought guidance or clarification. Any generated content has been fully understood and appropriately modified to align with the assignment.
2. We understand the submitted code and can explain our work if asked.

We declare that we have read, understood, and agree to abide this honor code.

Student names (of all the group members):

Student numbers (of all the group members):

Date:

Part A: Video Making (20 points)

In this part, your group will create a video exploring a specific topic related to Big Data or Data Analytics. The aim of the video is to demonstrate your group's understanding to effectively explain and showcase the topic that you have chosen. You can choose one of the following three topics:

1. Explain a Big Data tool/architecture

Choose a popular big data tool or architecture (such as Hadoop, Spark, NoSQL, Flink, Storm, Kafka, Hive, Tensorflow, etc). The tool/architecture can be related to data ingestion, data integration, data storage, etc. Your video should:

- Describe the purpose of the tool.
- Illustrate how the tool or the architecture works, and its primary functionalities.
- Provide a short demonstration showing the application of the tool.

2. Explain a Machine Learning model

Choose a machine learning model that has not been covered in the course. Here are some of the models that you can choose from:

Category	**Models**
Supervised Learning	Gradient Boosting, Naïve Bayes Classifier, Lasso Regression, Neural Nets, etc.
Unsupervised Learning	Hierarchical Clustering, LDA, Self-Organizing Maps, etc.
Other Models	Q-Learning, CNN, RNN, GAN, etc.
Transformer Models	BERT, GPT, etc.

Your video should:

- Describe the mechanics of the model and its underlying mathematical or algorithmic principles (briefly).
- Discuss applications of the model and where it is typically used.
- If applicable, implement the model in R/Python on a dataset.

3. A real-world Data Analytics case study

Research a real-world case study where data analytics was applied to solve a business problem. Your video should:

- Explain the context of the case study, including the industry, organization, and problem.
- Describe in detail how the data analytics project was implemented.
- Summarize lessons learned from the case study and its broader implications.

Note: If you have any other ideas (apart from the above three) that you believe would be valuable to explore, please check with the instructor for approval before proceeding.

Guideline for making video:

- Duration of the video: 8 to 10 minutes.
- Any number of members from your group can participate in creating the video.
- The target audience is your peers (fellow students). The video should be designed to communicate effectively with your peers. Make sure the content is relatable and presented in a way that your audience can connect with and understand.
- Simplify complex concepts and use examples or analogies where appropriate.
- Use clear visuals, diagrams, and animations to make the content engaging. Avoid heavy text slides.

Deliverable for Part A:

- An 8 to 10 minute video. Upload the video on Panopto/YouTube and share the video (or link to the video) on Canvas.

Part 2: Data Analytics (20 Points)

a) Business Context

EuroBank International (EBI) is facing the challenge of customer churn, which means that its customers are leaving their service for various reasons. The bank seeks to understand the underlying causes of customer attrition and proactively identify customers who are at risk of leaving. EBI has engaged your group to conduct a comprehensive data analytics study. Your task is to deliver useful insights to the bank.

b) Data dictionary

There are two datasets: “ebi_base_customers.csv” and “ebi_exp_customers.csv”. The two datasets have the same set of variables, except that the former dataset has churn information, whereas the latter dataset does not have that information. In parts c, d, and e below, please use the dataset “ebi_base_customers.csv”, and for part f, use the dataset “ebi_exp_customers.csv”.

Variable	Description
customer_id	Unique customer identifier
credit_score	Credit score of the customer
country	Country of residence of the customer
gender	Gender of the customer
age	Age of the customer
tenure	# of years the customer is having an account in the bank
balance	Account balance of the customer
products_number	Banking product purchased by the customer
credit_card	Does the customer have credit card? Yes: 1, No: 0
active_member	Is the customer an active member of the bank? Yes: 1, No: 0
estimated_salary	Estimated salary of the customer
churn	Churn status of the customer. Churn: 1, No churn: 0

c) Data pre-processing

Data pre-processing is a critical step in the data analysis process. This ensures accuracy and reliability of your analysis The goal here is to remove any errors or inconsistencies in the data and to transform the data into a suitable format for analysis.

- Are there any outliers/anomaly in the data that can distort the results? Address the outliers appropriately.
- Are there variables in the dataset that are not relevant to the analysis? Remove them.
- Are there categorical variables in the dataset? Consider encoding them to numerical

values if they are essential for your analysis.

d) Exploratory data analysis

The next step is to analyze the customer dataset to identify patterns and trends that could be contributing to customer churn. Answer the questions below based on your analysis.

- What is the overall customer churn rate in the dataset?
- How does the rate of customer churn vary by demographic variables such as age, gender, etc.? How does it vary across the countries?
- Is there a relationship between tenure and churn?
- Report interesting patterns that you find in the dataset.

e) Model building

The next step is to develop a predictive model to identify customers who are at risk of churning. Use at least 3 machine learning models to predict customer churn and answer the following questions:

- Which variables are the strongest predictors of customer churn? How did you conclude that these are the strongest predictors?
- How do different model evaluation metrics (e.g., accuracy, precision, recall) vary for different models?
- Which model would you use for predicting customer churn and why?

f) Recommendations

After developing the predictive model, the next step is to use it to identify customers who are at risk of churning. The bank can then take proactive measures to retain these customers, such as offering incentives, personalized services, or targeted marketing campaigns.

- Based on your analysis and domain knowledge, develop 3 recommendations that will help EBI to better manage customer churn. Explain the rationale behind those 3 recommendations.
- EBI has formulated a list of a subset of its current customers (see the dataset “ebi_exp_customers.csv” to answer this question) and would like to use your prediction model to take proactive measures to retain these customers. Specifically, the bank would target the customers (say, via telemarketing) who have high likelihood of churn.
  - Use your prediction model (from part e) to predict the likelihood of churn for each customer in the dataset “ebi_exp_customers.csv”.
  - Suppose that the value of retaining a customer is €5 while the cost incurred by the bank to avoid a customer from churning is €1. How many and which

customers from the dataset (“ebi_exp_customers.csv”) would you recommend the bank to target to maximize the total expected profit from this proactive targeting experiment? How would your answer change if the value of retaining a customer goes up to €10? Explain your computation.

g) Presentation Deck

After conducting the analysis, you need to present your analysis to the executive management of EBI. While conducting data analysis is important, communicating the results to the stakeholders is also important. Create a slide deck that illustrates the insights from data, your analysis, and recommendations. Few pointers to create the deck:

- Use visual aids such as charts, graphs, and tables to effectively communicate.
- Avoid technical jargon.
- Minimal use of long, verbose sentences.
- Be creative!

Deliverables for Part B:

- Project report – maximum 10 pages excluding appendices and references – that includes details of your analysis and answers to the questions. Remember to include the honor code on the cover page.
- Code file that includes the code to your analysis with appropriate comments. The project would be considered incomplete if you do not submit the code file that has all the analysis.
- Presentation deck – maximum 10 slides.

Grading rubric:

		Performance			
	Criteria	**Excellent**	**Satisfactory**	**Needs improvement**	**Incomplete**
Part A	Topic chosen, difficulty level, clarity and depth of explanation	10	8	3	0
	Accuracy of content	4	3	1	0
	Demonstration quality	3	2	1	0
	Organization, structure, audience connection	3	2	1	0
Part B	Exploratory data analysis	4	3	1	0
	Data pre-processing	3	2	1	0
	Model building	7	5	3	0
	Recommendation	3	2	1	0
	Presentation deck	3	2	1	0