Group Project

Maximum points: 40 Project report due date: 28th Nov. 2025 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 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 progressively penalized.
- Points Distribution: The distribution of points and the 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 by 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 that explores a specific topic related to Data Analytics. The goal is to demonstrate your understanding by clearly explaining and showcasing the chosen topic. You may select one of the following four themes:

1. Explain a Big Data tool/architecture

Select a well-known Big Data tool or architecture (see the table below for suggested examples). The chosen system may focus on data ingestion, integration, storage, or related processes. Your view should:

- Clearly describe the purpose and key features of the tool or architecture.
- Explain how it works, highlighting its main components and core functionalities.
- Include a brief demonstration that shows the tool in action.

2. Explain a Machine Learning model

Select a machine learning model that has not been covered in the course (see Table 1 for suggested examples). Your video should:

- Briefly explain how the model works, including its key mathematical or algorithmic principles.
- Discuss typical applications of the model and the types of problems it addresses.
- If applicable, demonstrate the model in R/Python using a suitable dataset.

3. Data-driven storytelling

Develop an engaging narrative that uses data to illustrate and communicate meaningful insights about a real-world phenomenon. The focus should be on turning data into a clear, compelling story rather than on technical modeling. Your video should:

- Identify a topic/question and gather relevant data (or use publicly available data) to explore it.
- Analyze and visualize the data to highlight patterns, trends, or relationships.
- Use visuals, infographics, or dashboards to make the story accessible and engaging.

4. A real-world Data Analytics case study

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

- Explain the context of the case study, including the industry, organization, and the

problem being solved.

- Describe in detail how the data analytics project was carried out. Include key methods, tools, and steps involved.
- Summarize the lessons learned from the case study and its broader implications.

The topics listed in the table below are only suggestive. You are free to choose any other relevant tool, model, or theme related to data analytics. Once you finalize the topic of your video, please register your topic here: Project Topic (column B).

Table 1: Suggested topics for Part A

Big Data Tools / Architectures	Machine Learning Models	Data-Driven Storytelling
Apache Hadoop	Gradient Boosting	Climate change trends
Apache Spark	Naïve Bayes Classifier	Smart cities
Apache Kafka	Lasso Regression	Financial markets
Apache Flink	Self-Organizing Maps	Music, pop culture
Apache Airflow	Hierarchical Clustering	Geo-politics
Apache NiFi	Latent Dirichlet Allocation	News dissemination
Apache Hive	Recurrent Neural Networks	Energy
Apache Storm	Convolutional Neural Networks	Computer vision
Apache Cassandra	Generative Adversarial Networks	Deepfakes
Apache HBase	Q-Learning	Robotics
MongoDB	Autoencoders	Payment systems
Elasticsearch	Transformer Architecture	Online misinformation
Snowflake	BERT	Social media sentiments
Databricks	Graph Neural Networks	Socio-economic patterns
Delta Lake	Explainable AI (SHAP, LIME, etc.)	Rise of AI, LLMs
…	…	…

Guidelines 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.
- Note: AI tools may be used for background research or planning purposes, but not for generating the video content itself. Videos found to be produced using AI will not be

graded.

Deliverable for Part A:

- An 8 to 10 minute video. Upload the video to Panopto/YouTube and share the link to the video on Canvas.
- Please also upload the link to your video here: Project Topic (column C). This will help other students view and learn the topic you chose.

Part B: 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	Number of banking products held by the customer
credit_card	Does the customer have any credit card with the bank? Yes: 1, No: 0
active_member	Is the customer an active member of the bank? Yes: 1, No: 0
estimated_salary	Estimated income 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 the 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/anomalies 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 into numerical values if they are essential for your analysis.

Document all pre-processing steps (e.g., handling missing values, scaling, encoding) clearly in your report and code file.

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. A few pointers to create the deck:

- The deck should summarize, not repeat, the report — focus on insights, visuals, and key takeaways.
- 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:

	Criteria	Performance			
		Exceptional!	Excellent	Satisfactory	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