Big Data 2026: Group project

Objectives:

As part of the evaluation of this course, students execute a Big Data project in Spark as a team to reach the learning objectives of solving and presenting an end-to-end solution to a Big Data problem in an intercultural team; and of demonstrating an expertise on key concepts, techniques, and trends (among others). In this project, they will apply the knowledge and techniques seen in class to a lifelike Big Data project. Furthermore, they will learn how to work in an intercultural team, how to develop and share business insights with a business audience, and technical analyses with a data (science) audience.

The format of this assignment is that you learn how to work on a Big Data problem independently. You help each other in your own team and are jointly responsible for the process and result. Based on your courses in the first semester and the materials seen in the Big Data course, you develop a project and data strategy to solve the various questions in this assignment and deal with the challenges you will be faced with in a logical way. By solving problems in a team, you learn how to overcome them and significantly improve your learning of solving Big Data projects using Spark!

Context:

Vauban 50 (or V-50) is a premium accessories brand founded in Lille in 2022 by Clémence, Mathis, and Santi, three former IESEG students who started working together as a team on a capstone project for a Digital Marketing course. Inspired by the architecture of the star-shaped Citadel de Lille and its iconic architect (the Marquis of Vauban), and referencing Lille's position on the 50th parallel, their brand Vauban 50 (or V-50) specializes in developing premium "personal architecture" accessories for urban professionals. Their key target audience is 16–34-year-olds with an active lifestyle who are looking for a local, sustainable, and high-performance product. Vauban 50 sells their backpacks exclusively online via their own website. They offer free shipping to customers for all their products, which incurs a cost of about €10 per order for the company, thanks to a favorable negotiated deal with a shipping company that handles all their shipping to customers.

They launched their first backpack, the CorePack, in 2022 with 3 new backpack models following over the next two years. Currently, they offer four products:

1. **CorePack**: Water-resistant, with one laptop sleeve and one main compartment, minimalist in slate grey.
2. **TechFortress**: Same exterior as the CorePack, but with added crush-proof padding, an integrated USB-C charging port, and a hidden anti-theft pocket, minimalist in matte black.

3. **AirLite**: Super light, minimal flat lay, uses a thinner, ultra-lightweight nylon, in electric blue. It has no internal organization, designed for those who just want a light bag to exercise or go on a short trip.
4. **EcoShell**: Designed to be easily foldable and tucked away, it is made from 100% recycled PET material (plastic bottles), in earthy green with nature accents.

Assignment:

V-50 hired your team as a business and data science consultant. After strategic meetings with the founders, it was decided that they want your team to identify and **predict the probability that a user visiting their website will order a backpack (1=order; 0=no order)**. They want to use this predictive model to identify if they should offer some incentives while customers are shopping to move them more fluently through the sales funnel and increase the likelihood of an order being placed.

In addition to focusing on prediction, also provide insights on **which criteria are important for increasing orders**. Think of **2 creative ways** (e.g., website features, marketing ideas) on how V-50 could improve this based on insights from the data. Describe these elements in the

business section of your presentation. This section should be written for middle-to- senior management responsible for business development.

Santi, the data scientist of V-50, has provided you with a training dataset of 5 tables of more than 28k orders placed between March 2022 and January 2025, in addition to 2 holdout datasets of website visits from February to March 2025. Your goal is to build a prediction model with the highest possible performance using the provided data, while respecting the fundamental principles of good data science. You can be creative and innovative how you use the available information (e.g., create new variables, merge tables), as you would do in practice! **The team that achieves the highest accuracy on the holdout dataset** for predicting which user will order a backpack, **gets +2 bonus points** on their assignment; respecting of course the appropriate modeling setup process and applying ethical practices (e.g., no AUC-hacking, data leakage, or other forms of cheating)! Furthermore, each team that, in addition to predicting which user will order a backpack, can **predict <u>which</u> product the user will order using a multi-class classification model** where, for each product, 1 means the specific product is ordered and 0 means the product (or any product) is not ordered, **gets +1 bonus point** (only if the code to do so is added to the notebook and the predicted score of 0 or 1 is added for each product to the .csv file; see deliverables). Describe your approach in the technical section of the presentation. This section should be concise and destined for a data science audience (e.g., describe variable creation, algorithms used, evaluation metric(s)).

Clémence, the marketing and business developer of V-50, has invested significantly in getting more traffic to their website using search ads (i.e., GSearch, BSearch) and, more recently, social ads (i.e., Social). She wants you to analyze **which channel/type leads to the most visits to their website and which is the most profitable**, so that she knows which channel(s)/type(s) to continue investing in in the future. In addition, she has conducted several experiments over the last few years to optimize the conversion rate of the website's home page. More specifically, she ran 5 different landing page designs (cfr. "/lander-1" to "/lander-5" in the column *pageview_url* in *website_pageviews*) as an alternative to the standard home page (cfr. "/home" in the column *pageview_url* in *website_pageviews*) to see how different landing page designs can improve conversions. **Analyze the performance of the different landing pages and provide a recommendation which she should continue using in the future.**

She has provided you with the following information on V-50's weighted average cost-per-click (CPC) across the used platforms:

Platform	**Type**	**2022**	**2023**	**2024**	**2025**
GSearch	Brand	€0.20	€0.25	€0.30	€0.35
BSearch	Brand	€0.10	€0.15	€0.20	€0.25
GSearch	Non-Brand	€0.80	€0.90	€1.25	€1.50
BSearch	Non-Brand	€0.40	€0.60	€0.80	€1.00
Social				€0.70	€0.80

Intermediate meetings:

During **Session 8**, a **Q&A session** will be organized per team where each team has 10 minutes to ask questions to the client. During this session, I will not act as your "professor" but a representative of the founders. Prepare your questions and prioritize them. You are hired as a professional Big Data science consultancy team so you will not get answers to questions like 'how do I handle missing values?', 'how should we deal with this variable?', 'tell me what I should do' etc.

Project organization:

This group project is organized in groups of 3 people. You can find the groups and their composition of team members under the Group project section on MyCourses.

Deliverables:

By **Wednesday, February 11th, 2026 (10:00)** upload the following items on the "Dropbox Deliverables" under the Group project section on MyCourses:

- A **.pdf presentation** of **max. 15-min** to be presented on Thursday, February 12th during class with 1/ an **executive management summary** (max. 2 slides) targeted at the senior management team of the company, in which you summarize the project's setup, main conclusions, and proposed actions + 2/ a **business section** focused on data understanding targeted at the business development team in which you discuss the relevant business insights and proposed business actions (max. 4 slides, incl. figures/tables) + 3/ a **technical section** (max. 4 slides, incl. figures/tables) targeted at the data science team in which you explain the data analysis strategy and methodology (e.g., feature engineering, algorithms used, model selection and performance) and your reasoning in more detail. It's recommended to use academic articles or other secondary materials to motivate your approach and findings. You can add **max. 5 slides to the appendix** at the end of your slide deck to go further in detail on your approach. **Think about visualization and write in a dense and concise way!** Every chart must serve a purpose, and every claim must be backed by data! Use slide titles as summaries for your main points and add key takeaways on the slides. **Everything should be professional, high quality and clear.** Name this **.pdf file** in the following format: *"BD26_slides_NameMember1_NameMember2_NameMember3.pdf"*

- A **.ipynb notebook** created in Colab using Spark with all code to execute the assignment. For this notebook:
  - In the first cell, write your team members' names, the academic year, and the course name. In the second cell, provide a path variable from where you read the datasets from.

- Write fast, efficient, and well-structured Spark code statements. Use Spark's functionalities as much as possible (e.g., pipeline(s), model building, tuning)!
- Act as a professional Big Data scientist and document your code well. Make sure to stick to the checklist for coding best practices on https://www.topcoder.com/coding-best-practices/. The idea is that the code you submitted is at the level for a data scientist at the company to take your notebook and easily run it and understand the code.
- As a reminder, you can download a notebook from Colab using "File" -> "Download" -> "Download as .ipynb". Name this **.ipynb file** in the following format: *"BD26_code_NameMember1_NameMember2_NameMember3.ipynb"*

- A **.csv file** consisting of two columns: *user_id* and *pred_score,* where the predicted score if an order is placed or not, is given per user_id using the holdout data (see folder "Holdout Data" on MyCourses). Note that you should not provide a probability but a score (1=order; 0=no order) for the column *pred_score*. If, in addition, you use a multi-class setup where you predict which specific product (i.e., product 1, 2, 3 or 4) will be ordered, your .csv file should also contain 4 columns where you give a score for each product (1=order; 0=no order) with column names *pred_multi_score_1* to *pred_multi_score_4*, where 1-4 represents the product_id as in the products table. Name this **.csv file** in the following format: *"BD26_pred_NameMember1_NameMember2_NameMember3.csv"*

Requirements:

- All code must be written in (Py)Spark utilizing Spark's methods, functions, and operations as much as possible.
- Use at least two algorithms in your modeling phase per model.
- Use at least two performance metrics per model.

Data description:

You can download the datasets under "Group project" on MyCourses.

In the table below you can find a description of the dimensions of the datasets:

Dataset	N° rows (sample)	N° rows (holdout)	N° columns
Products	4		3
Orders	28,991		8
Order_items	35,626		7
Website_sessions	434,008	38,861	10
Website_pageviews	1,052,523	103,278	4

In the next few tables, you can find a metadata description per dataset:

Products	
product_id	Unique identifier for the product (PK)
created_at	Timestamp when the product was launched
product_name	Name of the product

Orders	
order_id	Unique identifier for each order (PK)
created_at	Timestamp when the order was placed
website_session_id	Unique identifier for the website session (FK)
user_id	Unique identifier for the user (FK)
primary_product_id	Unique identifier for the primary product in the order if part of a bundle (FK)
items_purchased	Number of items in the order
price_euro	Total price for the items in the order
cogs_euro	Cost of goods sold for the items in the order

Order_items	
order_item_id	Unique identifier for each order item (PK)
created_at	Timestamp when the order was placed
order_id	Unique identifier for the order the item belongs to (FK)
product_id	Unique identifier for the product (FK)
is_primary_item	Binary variable with a value of 1 if it's the primary item in the order
price_euro	Price of the product
cogs_euro	Cost of goods sold of the product

Website_sessions	
website_session_id	Unique identifier for each website session (PK)
created_at	Timestamp when the session started
user_id	Unique identifier for the user (FK)
is_repeat_session	Binary variable with a value of 1 if the user has had a previous session
utm_source	UTM source parameter (traffic origin)
utm_campaign	UTM campaign parameter (marketing campaign name)
utm_content	UTM content parameter (ad/content variant)
device_type	Device category (mobile or desktop)
http_referer	URL for the UTM source
traffic_source	Type of traffic (direct, paid search, organic search, or paid social)

Website_pageviews	
website_pageview_id	Unique identifier for each website pageview (PK)
created_at	Timestamp for the pageview
website_session_id	Unique identifier for the website session the pageview belongs to (FK)
pageview_url	URL path for the pageview