Group project Big Data Tools 2024

Prof. Dr. S. Hoornaert

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 to a business team, and technical analyses to a data science team.

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 BDT, you develop a strategy how you can solve the various 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 (and each other)!

Context:

BLU is a French e-commerce player offering B2C and B2B customers a broad range of products across more than 100 product categories ranging from kitchen utensils to computer games. They have been active for 5 years and achieved a nice growth of 16% in 2020, 12% in 2021, and 8% in 2022. To get more insights how the company is performing compared to its competitors, the marketing team ordered a report from a consumer insights agency comparing BLU to its two main competitors Amazon.fr and Cdiscount.fr. This revealed that BLU is able to acquire a larger share of new customers compared to the other two (7.8% versus 2.8% for Amazon and 3.4% for Cdiscount) but is underperforming in repeat business from existing customers (6.2% vs. 19.6% for Amazon and 14.4% for Cdiscount). As a next step, BLU’s marketing team wants to have a deeper understanding of its customer base and use predictive modeling to inform its business decisions.

Assignment:

BLU hired your team as a data science consultant. After strategic meetings with the marketing and data science teams, it was decided they want your team to identify and predict what factors lead some customers to give a positive or negative review score on an order. Whenever a customer places an order, their marketing automation software automatically sends an email after a certain amount of time with a link to a customer satisfaction survey. This survey traditionally gets a very high response rate (due to some incentives like vouchers).

They want you to make a prediction model to predict whether the review score of a given order will be positive (4-5 on 5) or negative (1-3 on 5) given some input features.

BLU's data science team provided a dataset of 6 tables of +/- 50k orders placed between September 2020 and June 2022. 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, use unstructured content, etc); as you would do in practice! The team that achieves the highest accuracy on the hold-out sample gets +2 bonus points on their assignment; respecting of course the appropriate modelling setup process and applying ethical practices (e.g., no AUC-hacking or other methods of cheating)! Furthermore, each team that solves the case using a multi-class classification model (where a probability is given per label) gets +1 bonus point. 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, cross-validation approach, evaluation metric(s)).

In addition to focusing on prediction, also provide insights on which criteria are important for improving review scores. Think of 2 creative ways (e.g., website features, marketing efforts) on how BLU could improve this based on the 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.

Intermediate meetings:

During **Session 8 (Feb. 1st)**, a **Q&A session** will be organized per team where each team has 10-15 min. to ask questions to the client. During this session, I will not act as "your professor" but a representative of the client. Prepare your questions and prioritize them. You are hired as a professional 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. The groups were assigned at random. You can find the groups and their composition of team members on MyCourses.

Deliverables:

By **Wednesday, February 14th (10:00)** upload the following items on the deliverables dropbox under the 'Group project' section on MyCourses:

- A **presentation** of **max. 15-min** to be presented on Thursday, February 15th during class with 1/ an executive management summary (max. 2 slides) targeted at the middle/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 insights and 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., definition of variables, variables included in the model, algorithms used, cross-validation technique) 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 at the end of your slide deck in appendix to go further in detail on your approach. Think about visualization and write in a dense and concise way. Use slide titles as summaries for your main points. Everything should be professional, qualitative and clear.

- A **notebook** created in Databricks 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 in the datasets.
  - 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/.
  - As a reminder, you can download your notebook from Databricks using "File" -> "Export" -> "Source File". Name this .py file in the following format: "BDT_2024_LastNameTeamMember1_LastNameTeamMember2_LastNameTeamMember3.py"

- An **Excel file** consisting of two columns: *order_id* and *pred_review_score*; where the predicted review score is given per order_id in the test set for the July – September 2022 period (see folder "Holdout data" on MyCourses). Name this Excel file as "BDT_2024_LastNameTeamMember1_LastNameTeamMember2_LastNameTeamMember3.xlsx"

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 from MyCourses. In the table below you can find the dimensions of the datasets.

Dataset	**Number of rows**	**Number of columns**
products	20,500	9
orders	51,449	8
order_items	58,501	5
order_payments	53,847	5
order_reviews	51,433	5

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

Products	
product_id	Product unique identifier
product_name_length	Number of characters in the product name
product_description_length	Number of characters in the product description
product_photos_qty	Number of photos included in the product description
product_weight_g	Product weight (in grams)
product_length_cm	Product dimensions - length (in centimeters)
product_height_cm	Product dimensions - height (in centimeters)
product_width_cm	Product dimensions - width (in centimeters)
product_category_name	Product category name

Orders	
order_id	Order unique identifier
customer_id	Unique identifier of the customer
order_status	Status of the order
order_purchase_timestamp	Timestamp at which the customer purchased the order
order_approved_at	Timestamp at which the order was verified and approved by the company
order_delivered_carrier_date	Timestamp at which the order was delivered by the company to the logistic partner
order_delivered_customer_date	Timestamp at which the order was delivered to the customer by the logistic partner
order_estimated_delivery_date	Estimated delivery date that was communicated to the customer at the purchase moment

Order items	
order_id	Order unique identifier
order_item_id	Sequential number identifying the order of the ordered items. A customer can order multiple items per order.
product_id	Product unique identifier
price	Item price in euro (excl. VAT)
shipping_cost	Cost for shipping the item to the customer in euro (excl. VAT)

Order payments	
order_id	Order unique identifier
payment_sequential	Sequential number identifying the order of the payment types used. A customer may use several payment types for one order.
payment_type	Method of payment chosen by the customer
payment_installments	Number of installments chosen by the customer
payment_value	Order value in euro

Order reviews	
review_id	Unique identifier for a review
order_id	Order unique identifier
review_score	Score from 1 to 5 given by the customer on a customer satisfaction survey
review_creation_date	Date at which the customer satisfaction survey was sent to the customer.
review_answer_timestamp	Timestamp at which the customer answered to the customer satisfaction survey.