END-TO-END MACHINE LEARNING

DSMarket

Luis Roque
Tiago Otto
João Reis

Based on the work of Pedro Costa and Filipe Calderero

NUCLIO
DIGITAL SCHOOL

DSMarket - your next generation store

Welcome to the first graded task in the International Master in Data Science: the **DSMarket** case!

The DSMarket case is presented as a **role play practical exercise**, divided in several tasks that you will have to complete during the following months, and that you will be asked to submit and present at the end of the master.

This practical exercise aims to **recreate a realistic working scenario for a data scientist**. The success of the different projects will often depend on the combination of the three main types of skills that we have already talked so much about (**programming + analytic + business**). The expected approaches to follow for each task are often not specified, and their requirements won't be always 100% clear (welcome to Data Science uncertainty!)

This project will also provide an opportunity to **work in groups**, to work with one another's codes, and to have your first exposure to the collaborative tools that are frequently used in almost every DS project.

You are about to become Nicole, a **Senior Data Scientist** joining the financial department of a small chain of shopping centres: DSMarket.

Have fun!

Context

DSMarket (previously known as *TradiStores*) is a small chain of shopping centres across the USA that has found itself amongst the very latecomers to the digital transformation that has been reshaping the retail sector for quite a few years. Their change of name is only the first step of a disruptive five years plan to completely remodel each single process within the company. DSMarket has incorporated Michelle Huggins as their new Chief Digital Officer. After more than 15 years of experience leading the Digital Marketing Areas of key companies within the retail sector, Michelle is surely planning to exploit the key asset that DSMarket has been accumulating but ignoring for too long: ***its data!***

Along with the many digital marketing specialists that Michelle is hiring during the first year, she has only included **one data scientist**. Data initiatives during the first year will focus in the standardization and transformation of the company's data sources, and in the migration of all sources and data processes to the cloud. Data engineers and data architects will be the main tech profiles required. DSMarket is nonetheless interested in incorporating a senior data scientist to boost the DS initiatives with higher priority. The DS team is planned to rapidly increase from the second year.

You will be Nicole. Nicole has been hired as a **senior data scientist** by the new Chief Digital Officer. However, you will be **directly reporting to Paul Rogers, the Finance Director of the company**. The initial DS initiatives that have been prioritized are issues of greatest importance for the financial department. Sales predictions in DSMarket have been always done using very rudimentary approaches, and the margins of error obtained are affecting many areas of the company. The magnitude of those errors stopped being acceptable a very long time ago. In addition, many of the internal processes within the company (stock estimations, prices optimization, deliveries, stockout predictions, …) are very manual processes with a strong dependence on business experience, and their optimization using AI methods has been included as part of the 5 years plan that the company has drafted.

Google
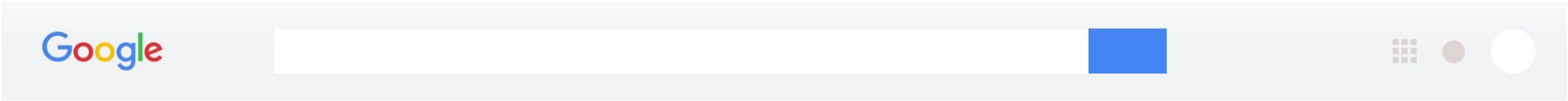
From: Justin (HR Director)

September 20th, 2021
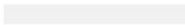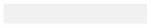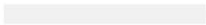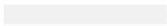
To: Nicole

Good morning Nicole!

Welcome to the DSMarket family.

Everyone is looking forward to finally meeting you, especially Michelle (our Chief Digital Officer). The early incorporation of a Data Scientist into the team has been one of her priorities. She considers you a key element to transform DSMarket into a "data driven" company. She has a lot of ideas for your initial months with us. I'm sure you will be hearing from her very soon.

As we discussed during the interview, you'll be directly reporting to Paul Rogers, the Finance Director of the company. That might change in the future, once the full DS team is established. Paul is currently out of the office, but he wanted to schedule a call to say hi at some point during the week.

I wish you a great start in the new job. If there is anything I can help with, please let me know.
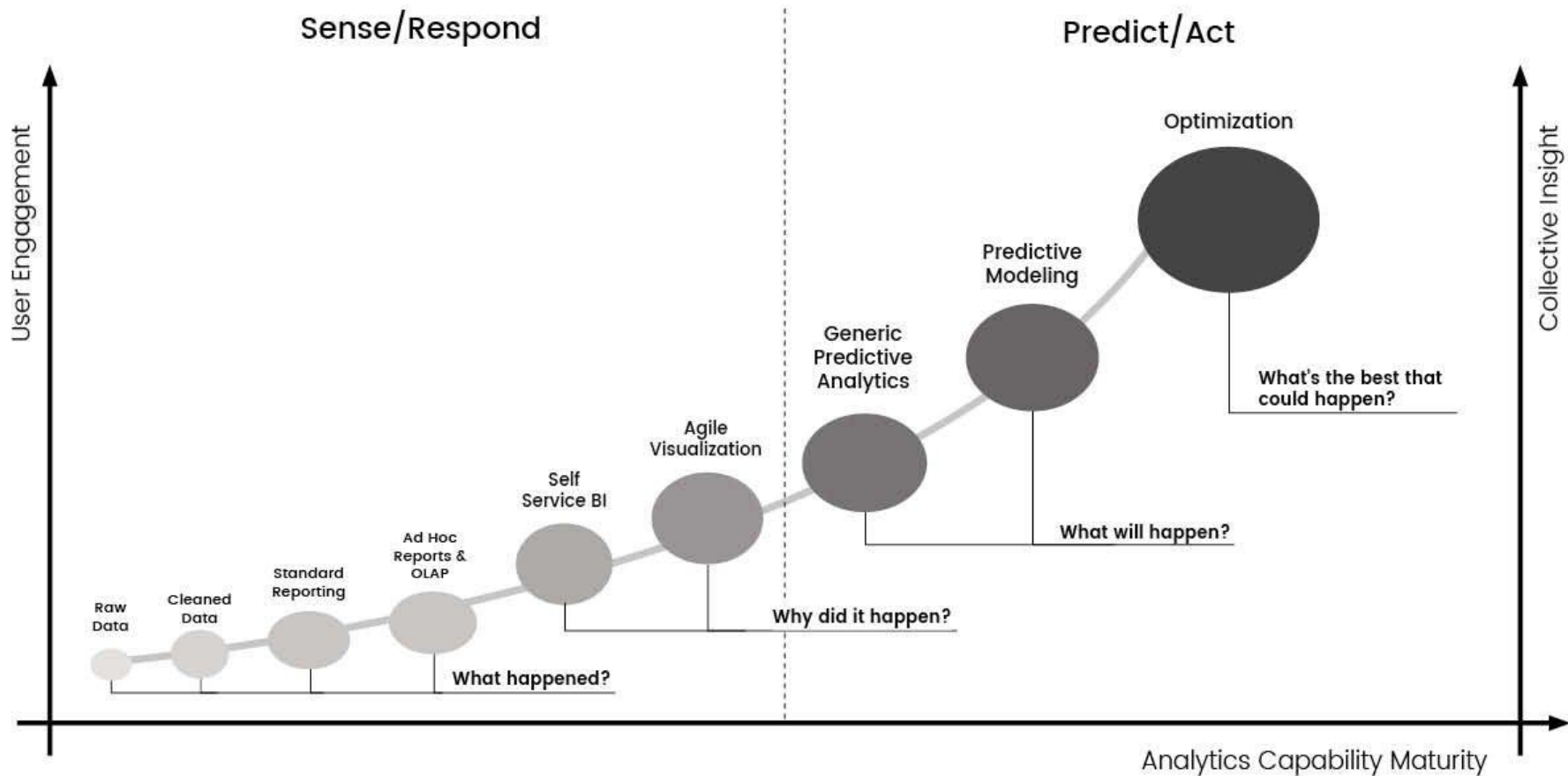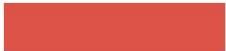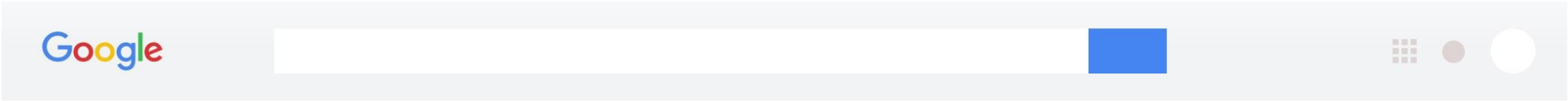
See you soon,

Justin

Google

From: Michelle (Chief Digital Officer)

To: Nicole

September 20th, 2021

Hello Nicole,

First of all, welcome to DSMarket!

I don't want to be the one overloading you with information on the first day. But once you are all set up, there are plenty of data science initiatives I'd like you to start with. DSMarket has always had a very "traditional" way of doing things, but it's never too late to join the digital transformation party. We might be at a very early stage in the analytics maturity curve, but with your help I'm sure we'll start progressing at a quicker pace.

I have already asked Martin (our Technology Lead) to give you access to a sample of the main data sources of the company. Your work during these f**irst months will focus in our stores from 3 big cities (NY, Boston & Philadelphia)**. Let's start testing your magic there, and we'll expand at a later stage to the rest of the country.

I'm sure your contribution will be one the main pillars for our journey. I'm very looking forward to start working with you.

M

Advanced Analytics Maturity Curve

Google

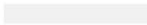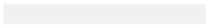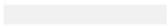**From:** Martin (Technology Lead)
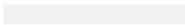
To: Nicole

Hey Nicole,

Welcome to the team.

Michelle has requested your access to some of the data sources of the company. We have decided to dump the necessary tables in shared folder for the time being. That should allow you to start working right away. I'm not particularly in favour of granting more permissions than necessary, and it seems that during your initial months you'll be focusing in a defined sample of stores. We will schedule a couple of training sessions on the information systems of the company for you, but I don't think that's a priority at the moment.

I've attached a file with all the relevant information you'll need to understand the tables in the shared folder (mainly column descriptions).

Best regards,

Martin

FiILE 1. daily_calendar_with_events.csv

| Name	Table	Description
date	calendar	date in y-m-d format
weekday	calendar	day of the week
weekday_int	calendar	numeric day of the week (Saturday day 1, Friday day 7)
d	calendar	day identifier
event	calendar	if the date includes an event, the name of this event (only a few are included)

FiILE 2. item_prices.csv

Name	Table	Description
item	prices	product id
category	prices	product category
store_code	prices	alphanumeric code of the store
yearweek	prices	date period for the price (year-week format)
sell_price	prices	price for the product "item" for the period in "yearweek". Prices are provided per week (average across 7 days). If not available, there were no sales for the product during that week

FiILE 3. item_sales.csv

Name	Table	Description
id	sales	sales series id (combination of item + store_code)
item	sales	product id
category	sales	product category
department	sales	department id (different identifier for different stores)
store	sales	store name
store_code	sales	store id
region	sales	region
d_1,d_2,d_...	sales	number of units sold per day

Task 1: Analysis
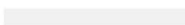
Google
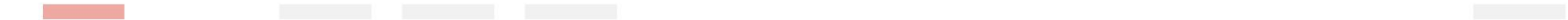
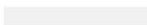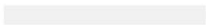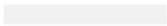**From:** Michelle (Chief Digital Officer)   **September 22th, 2021**
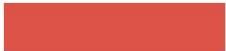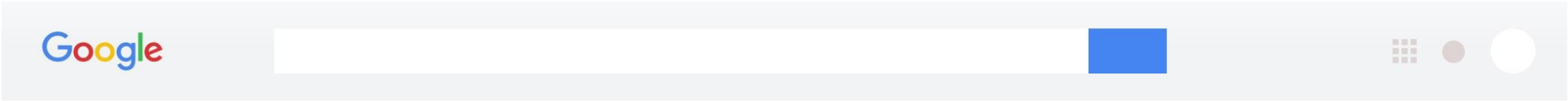
To: Nicole
CC: Paul (Finance Director)

Hi Nicole,

I hope you are settling in well in your new role. Martin has mentioned that you have already been able to access our databases. I'm happy to hear that!

Since I joined DSMarket, I've been wanting to analyse in depth the current picture of the company. So far I've been looking at global sales trends, but I really would like to **evaluate every angle of our activity.** I'd like you to help me with that. It would really appreciate if you could start looking at the data from **NY, Boston and Philly**. My intuition says that we probably have **some products that are not so popular any more**, and it's likely that **most popular products vary across cities**, or even across **stores** (which might vary in prices as well). Our marketing actions will be exploiting those differences. We need to understand every single detail of the business! I trust you for that 😊.

You and Paul should actually present your results to the executive board. What would be reasonable date to schedule that meeting? Thanks very much Nicole!!

M

Google

From: Michelle (Chief Digital Officer)

September 23th, 2021

To: Nicole
CC: Paul (Finance Director)
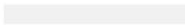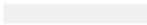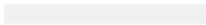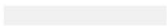
Hi Nicole,

Almost forgot. Do you think you could also work in a BI service that would allow us to follow the main results of your analysis on a regular basis? That would be super useful for the executive board. Just use the **dashboarding solution** that you feel more comfortable with.

Thanks again,

M

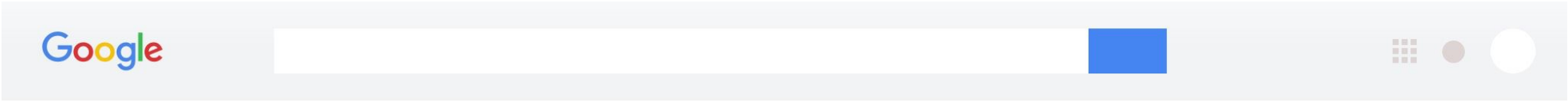# Task 2: Clustering

Google

From: Joelle (Marketing Manager)   **September 23th, 2021**

To: Nicole

CC: Paul (Finance Director), Michelle (Chief Digital Officer)

Hi Nicole,

First of all, welcome to the company!

Talking to Michelle, we've thought it would be a great idea if we could **identify groups of products that behave in a similar way**. Michelle was saying that with your magic it's easy to identify groups of similar products, and such groups will be super useful to evaluate the performance of our **different campaigns**. **How many groups do you think we should consider? 5? 10? 20?**

Also, do you think we could find a "solid" approach to identify how similar are stores to one another? Would store clustering also make sense here? Could you also do that?

I'm glad that we can finally count on someone with your skills within the team!

Best regards,

Joelle

Task 3: Sales Forecasting

Google

From: Paul (Finance Director)

To: Nicole

September 27th, 2021

Good morning Nicole,

I hope your first week went fine. I've seen that you have become quite popular already.

Now that you already have access to the company's databases, there is one particular initiative that I'd like you to focus in. DSMarket has been always depending on rudimentary approaches to **forecast product sales**. As we discussed during your interview, the main reason behind your early incorporation was the development of a **new forecasting approach**. We have an important number of use cases that do depend on the accuracy of those predictions. I am confident that with your lengthy experience with predictive models we can do a much better job than we are doing now.

We usually start predicting sales at a store-product level, and to obtain aggregated sales per department/store/city we add up the independent predictions. Would that be still a valid approach? **Let's start with 28 days predictions (4 weeks).**

Paul

Google

From: Martin (Technology Lead)

To: Paul (Finance Director)
CC: Nicole

September 28th, 2021

Hey Paul,

Yes, we will happy to participate in the evaluation of the sales prediction models. We have a few additional weeks of data available with which we will be able to test the predictive error of the model.

@Nicole, we will be sending you the output format that we will require to evaluate your models.

Thanks,

Martin

NUCLIO
DIGITAL SCHOOL

Task 4:
Store Replenishment
Use Case

DS MARKET
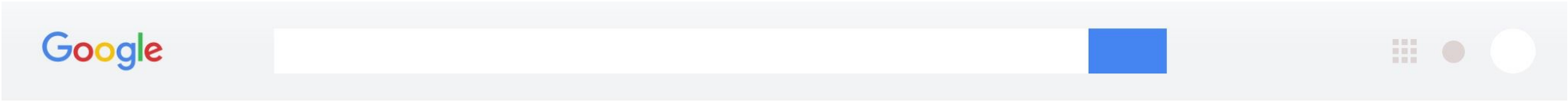
Google
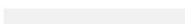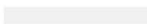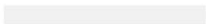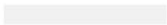
Hey Nicole,

How are those sales predictions coming up? The Operations Department is already considering the application of your advanced predictive models for one of their most critical use cases: **stores replenishment**. Albeit if needed, the stock for some products can be replenished on a daily basis, the supplies to provision stores are **mostly distributed on a weekly basis** (beginning of each week). You're probably already visualizing the importance of sales predictions for that. Minimizing the remanent stock is every retailer's desire, but that desire is even stronger for supermarket items.

Could you please draft a **proposal detailing your solution to apply the sales predictors to the stores supply use case?** We would also need to specify any extensions that the models might require, as well as the productization details for the use case. Martin requested the deployment of an API for its execution. Is it something you would be comfortable with? Do not worry about the implementation for now, let's just focus on that draft.

Thanks,

Paul
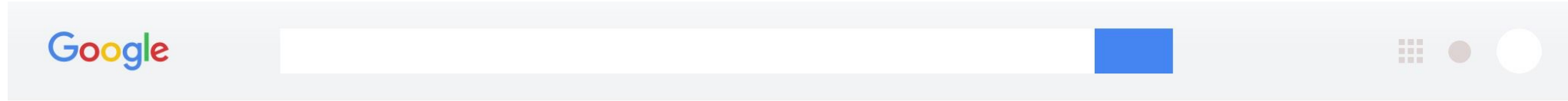
Google

From: Michelle (Chief Digital Officer)

To: Nicole

CC: Paul (Finance Director)

October 4th, 2021

Hi Nicole,

I have high expectations of the application of the sales prediction models to the stores supply problem. This a very well chosen use case to start demonstrating the benefits associated to the application of data science to every level of the company. I have no doubt that our solution based on ML algorithms will significantly improve the way in which DSMarket **has been estimating supplies per store all this time**.

Before going big across the country, I would like you to design a **pilot tes**t to confidently demonstrate the improvements directly associated to the new approach (i.e. dollars!). The creation of a data driven culture should be one of our priorities at this stage, and it is important that everyone starts believing in the power of this type of algorithms. I'm thinking of a case-control study (maybe by products or stores? maybe both?). You are the expert! What would you do and how would you evaluate it? Can you please send me a proposal?

Thank you for the hard work!

M

DSMarket - your next generation store

NUCLIO
DIGITAL SCHOOL

Rules of engagement:
- The assignment is to be carried out in groups of 2-3 people
- Communication channel between team members and tutor will be Slack
- It will be valued not only the technical level of the different tasks, but also the creativity, business orientation and the ability to communicate the results
- The ability to produce well-structured results and to follow clean coding principles will also be evaluated
- Expected output (deliverables) for the assignment are the following:
  1. Technical document with methodology and results (academic report)
  2. Requested deliverables for each independent task (dashboard, codes, outputs from the modelling in the required format, and requested proposals for each task)
  3. Final presentation for the executive board

Evaluation:
- Analysis 20%
- Clustering 20%
- Sales prediction model 30%
- Store replenishment use case (with MLOps) 20%
- Evaluation design 10%

Dates:
- October 20th - Deadline for reports and deliverables
- October 26th -Presentation to executive board