Data Scientist
Practical Exercise

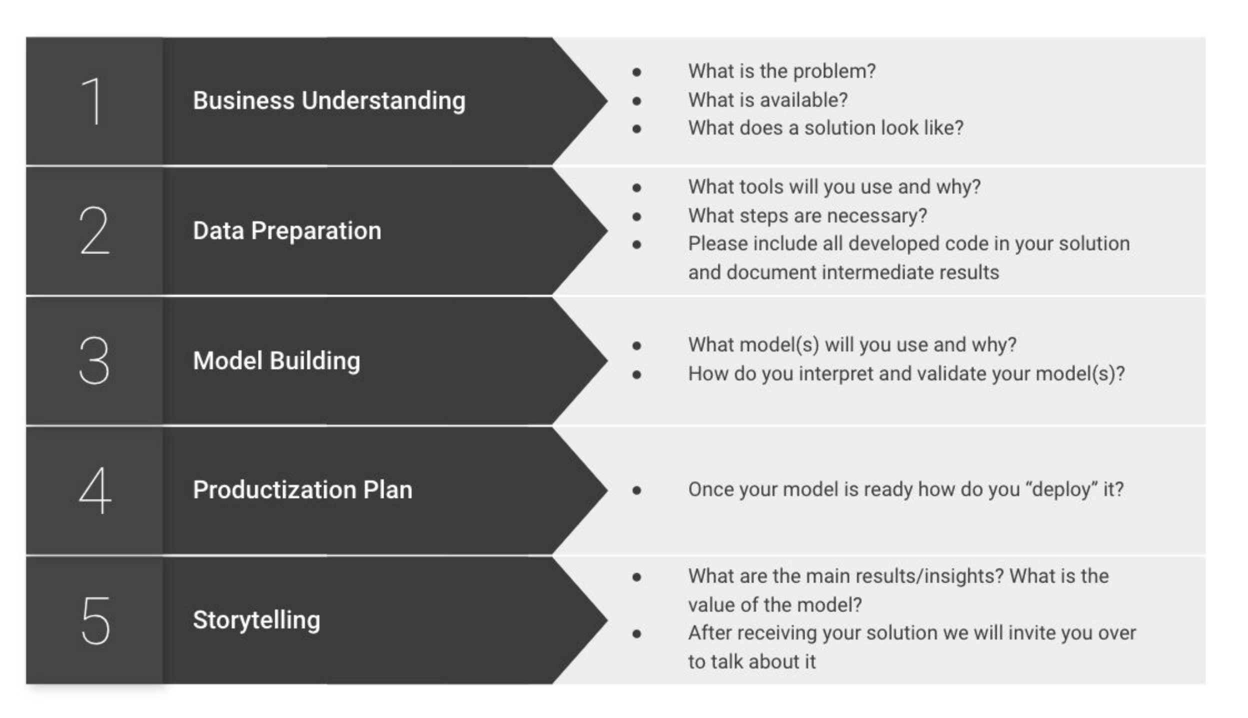
Introduction

These are the dimensions on which we will evaluate your work:

- Business understanding and storytelling
- Grasp of relevant methods and underlying concepts
- Critical thinking
- Technical acumen and tool proficiency

You are completely free to define your approach to the problem described below and add as much complexity to your solution as desired. Keep in mind however that we expect you to maximize the business value of your output vs the complexity of your approach at all times.

Please keep in mind the following milestones as part of your approach:



1	Business Understanding	• What is the problem? • What is available? • What does a solution look like?
2	Data Preparation	• What tools will you use and why? • What steps are necessary? • Please include all developed code in your solution and document intermediate results
3	Model Building	• What model(s) will you use and why? • How do you interpret and validate your model(s)?
4	Productization Plan	• Once your model is ready how do you "deploy" it?
5	Storytelling	• What are the main results/insights? What is the value of the model? • After receiving your solution we will invite you over to talk about it

Data Scientist

Practical Exercise

Part I - Conceptual Questions

1. Would you rather have too many false positives, or too many false negatives? Explain.
2. Give an example from your work experience of selection bias. Was it important? In general, how can you avoid it?
3. What is the difference between Bayesian Estimate and Maximum Likelihood Estimation (MLE)? Can you think of an example where Bayesian estimate is the most appropriate method? Another example in which MLE is the most appropriate method?

Part II - Conceptual Problem

The table shows data for two retail stores. It includes the average weekly sales, the standard deviation of the data and the number of data points. You are in charge of giving a bonus to the store managers based on performance. What would you do?

Downtown store	**Suburban store**
Average weekly sales = $800,000 std dev = $100,000 n_1= 50	Average weekly sales = $780,000 std dev = $30,000 n_2= 50

Part III - Practical Exercise

The data that will be provided to you comes (with modifications) from drivendata.org. They had a competition to predict from household surveys if a household is poor or not. You can find more context here.

Please use your machine learning skills and experience to develop a prediction model base on the data provided. We are not interested as much in the final result as in the process you follow, the insights you obtain, and how you communicate your results and findings.

Instructions & next steps

Within a week of receiving the exercise, you'll send us back your answers for I and II, and a report explaining the methods you used and your results, including code, plots and anything else you find useful. If your exercise is successful, you will be asked to present your results together with your previous work to us at our offices in Madrid.

Thanks in advance!