

## Oliver Wyman Online Case Study: Data Science Profile

### **Background:**

Pharmaceutical companies are evolving towards more data-driven decisions to build their Commercial and R&D strategies. One of the many places where analytics plays an important role is in harvesting this data to enable better planning and execution of Sales and Marketing propositions. One such use case is leveraging advanced algorithms to develop an in-depth understanding of global markets and disease areas, along with the factors that affect their forecasts and marketing plans

### **Problem Statement:**

The client needs to come up with a robust forecasting algorithm to predict sales for a portfolio of two drugs which may be impacted by market factors like account demographics, promotional expenses, product launch, etc.

To help with this problem, the information you will have available is:

1. Account level historical sales data for 8 years for both the products (Drug\_1 and Drug\_2)
2. Post launch Sales data for a third product in the market (Drug\_3)
3. Total Promotional Expenses and a split of Sales Force and Digital expenses of these products
4. Potential mapping for each account based on patient size – Low, Medium, High

The ask is for monthly sales to be **predicted for the next 9 months at the account level for each of the two drugs (Drug\_1, Drug\_2)** in Country 1, Country 2, Country 3, Country 4, and Country 5 using the available information

### **Data Available:**

**Account Sales and Expense File:** 8 years of sales data for Drug\_1 and Drug\_2 for 25 accounts in each country (1 to 5) along with monthly expenses, input file also includes post-launch sales data for an additional product in the market (Drug\_3)

(Dummy Data snapshot)

ACCOUNT_ID	PRODUCT	MONTH	COUNTRY	UNIT_SALES (in Million \$)	Total_Expense (in Million \$)	SalesForce_Expense (in Million \$)	Digital_Expense (in Million \$)
376321	Drug_1	8/1/2010	COUNTRY 1	55.6651672	0.521504	0.312902	0.204301
376321	Drug_1	9/1/2010	COUNTRY 1	42.98804927	0.809671	0.485803	0.261934
376321	Drug_1	10/1/2010	COUNTRY 1	26.49380457	1.112334	0.6674	0.322467
376321	Drug_3	11/1/2010	COUNTRY 1	68.09930294	0.985907	0.591544	0.297181
376321	Drug_2	12/1/2010	COUNTRY 1	34.03740076	1.080891	0.648534	0.316178
376321	Drug_2	1/1/2011	COUNTRY 1	36.13000728	1.126302	0.48123	0.626428

**Account Potential:** Flags to mark the potential of each account based on patient size- classified as high, medium, and low potential

(Dummy Data snapshot)

ACCOUNT_ID	ACCOUNT_POTENTIAL
3754	Low
2134	Low

7254	Medium
27174	High
77361	Medium
7462	Low
9361	High

### Submission Guidelines:

#### Codes:

- Submitted codes must be written in Python, along with a README file with a description of how to run the model to predict the dependent variable and generate the <userID>.csv file
- There is no limit on execution time, but the code should generate the output file:<UserID>.csv
- Please ensure that the submitted code files are well-commented and structured for easy comprehension.

#### Expected Output File:

- Upload a zip file (max allowed size is 5MB)
- Submission format: Please use the format shared with the case study to populate the monthly Sales forecast along with Lower and Upper Prediction intervals for Drug\_1 and Drug\_2 in 5 countries at the account level

ACCOUNT_ID	COUNTRY	PRODUCT	YEAR	MONTH	Sales Forecast (in Million \$)	Lower Prediction Interval (in Million \$)	Upper Prediction Interval (in Million \$)
376321	COUNTRY 1	Drug_1	2019	1			
2097539	COUNTRY 1	Drug_1	2019	1			
4613	COUNTRY 1	Drug_1	2019	1			
541958	COUNTRY 1	Drug_1	2019	1			
609418	COUNTRY 1	Drug_1	2019	1			
655243	COUNTRY 1	Drug_1	2019	1			

- Evaluation will be based on methodology and algorithm/ approach used to solve the problem in addition to quantitative assessment of the output file (sales and prediction intervals)
- Bonus points will be assigned for use of the XG boost algorithm for structuring and solving the problem

#### Methodology Description:

Candidates are expected to submit a presentation having the methodology used, learnings while running the process, and insights from data. Bonus points will be assigned for providing a brief on all other methodologies and algorithms that could be used to solve the problem along with their pros and cons.