Junior/Intern Data Scientist-Machine Learning Engineering Positions

Take-home test 1Q22

Deadline: Sunday, February 20th 2022 at 23:59 PM

Congratulations on reaching this stage of the recruiting process. This test consists of two sections. In the first one, you are required to use an analytical tool (Python, R, Excel, etc) to solve each of the questions. For the second section, you are required to choose one among three Machine Learning use cases and elaborate on it.
In this test, we will assess the following:

1. Creativity
2. Analytical thinking
3. Machine Learning solutions mindset
4. Data wrangling and exploration skills
5. Written communication skills
6. Data visualization skills

You must deliver a PDF file elaborating on each section and any additional resources to solve the questions. We also expect you to justify your answers. Please include lines of code, plots, diagrams, everything you need to communicate your solutions. Finally, create a zip file and send it back in the email received with the take-home test instructions.

1. Data Wrangling + Exploration

Set of 4 questions to be solved with two data sets [1]:

- Histórico de turismo en México (2016-2021). **File: turismo_mexico.csv**
- Pueblos mágicos de México. **File: pueblos_magicos.csv**

Questions:

1. **Exploratory Data Analysis**: What were the ten Pueblos mágicos with the most population in 2015? What were the ten Pueblos mágicos with the least population in 2010? *Data set: Pueblos mágicos de México.*
2. **Data Wrangling**: Change Mexico states full name for the three characters ISO convention. Store it in a variable called: estado_iso. After that, display the new variable's unique values. Use the following table as a reference. *Data set: Pueblos mágicos de México.*

Full name	3-letter-code (ISO code)

[1]Datatur, Análisis integral del turismo, http://www.datatur.sectur.gob.mx/

Aguascalientes	AGU
Baja California	BCN
Baja California Sur	BCS
Campeche	CAM
Coahuila	COA
Colima	COL
Chiapas	CHP
Chihuahua	CHH
Durango	DUR
Guanajuato	GUA
Guerrero	GRO
Hidalgo	HID
Jalisco	JAL
Mexico	MEX
Michoacan	MIC
Morelos	MOR
Nayarit	NAY
Nuevo Leon	NLE
Oaxaca	OAX
Puebla	PUE
Queretaro	QUE
Quintana Roo	ROO
San Luis Potosi	SLP
Sinaloa	SIN
Sonora	SON
Tabasco	TAB
Tamaulipas	TAM
Tlaxcala	TLA
Veracruz	VER
Yucatan	YUC
Zacatecas	ZAC

[1]Datatur, Análisis integral del turismo, http://www.datatur.sectur.gob.mx/

3. **Analysis**: The leadership team wants to review the historical evolution of International tourism in Mexico. Use the Histórico de turismo en México data set to elaborate an executive summary of a document that depicts the answer to the leadership team's request. Justify your answers. *Data set: Tourism in Mexico over time.*
4. **Creativity to communicate analytical results**: The Leadership team wants to know the number of Pueblos mágicos in each state. You, therefore, create an analysis. How would you communicate your results? **Important:** The leadership team has a non-technical background, so you might find data visualization techniques helpful. Elaborate a paragraph describing how to interpret your results and justify your answers. *Data set: Pueblos mágicos de México.*

2. Machine Learning solutions mindset

In this part, you are required to choose one of the following use cases. We encourage you to leverage data to solve it, including Machine Learning and business rules based on data analysis, but feel free to provide any analytical solution for it.

1. **Business use case 1:** Solution to accept credit card customers in a digital bank.
2. **Business use case 2:** Solution to match a customer's selfie with an ID card photo.
3. **Business use case 3:** Solution to classify social media comments sentiments (positive or negative).

After choosing one use case, we ask you to solve the following questions. Feel free to use diagrams or any additional material to complement your answers. *You are not required to code your solutions.*

- Describe your end-to-end approach to solve the problem.
- Describe in detail the data and statistical approach. **Hint**: Try to think of the data and analytical solution for the use case you chose.
- Describe in detail how you would deploy this solution in production so that it could be executed in real-time. **Hint**: Review MLOps, AWS, Google Cloud or Microsoft Azure microservices documentation.

Good luck!

[1]Datatur, Análisis integral del turismo, http://www.datatur.sectur.gob.mx/

Appendix

Data sets layout

- Histórico de turismo en México (2016-2021). **File: turismo_mexico.csv**

Variable	**Description**
fecha	Date (dd/mm/yy)
visitantes_internacionales	Total number of international visitors coming to Mexico (in Thousands)
turismo_al_interior	Number of international visitors coming to Mexico in the category of “Turismo al interior” (in Thousands)
turismo_fronterizo	Number of international visitors coming to Mexico in the category of “Turismo Fronterizo” (in Thousands)
excursionistas _fronterizos	Number of international visitors coming to Mexico in the category of “Excursionistas Fronterizos” (in Thousands)
pasajeros_crucero	Number of international visitors coming to Mexico in the category of “Pasajeros en Crucero” (in Thousands)

- Pueblos mágicos de México. **File: pueblos_magicos.csv**

Variable	**Description**
pueblo_magico	Pueblo mágico’s name
estado	Pueblo mágico’s state name
pob_2010	Number of habitants in a certain pueblo mágico in 2010
pob_2015	Number of habitants in a certain pueblo mágico in 2015

[1]Datatur, Análisis integral del turismo, http://www.datatur.sectur.gob.mx/