Data Science recruitment process

Technical test: Machine Learning – Operations

Context of the test

Pivot and Co is an AI and strategic advisory consultancy, with acceleration teams focused on data-driven growth & transformation in medium and large organizations. We are convinced that decisions enabled by data analytics can contribute to mobilize organizations and deeply change how people are operating and create sustainable impact.

In the data science team, we develop tools that allow analysing, modelling and adding intelligence to companies' proprietary datasets. One important step in this process is to collect, complete, normalize and perform statistical analyses of the data. Therefore, Python libraries such as pandas, numpy, matplotlib, seaborn and all other regular expression pattern libraries are essential for best results. Another step in the process is to select and design the most appropriate machine learning model to answer clients' need. For this step we rely mainly on sklearn, but we also use other specific libraries like fbprophet, tensorflow, keras, pytorch, … depending on the project. A last step in this process and on which we are putting more and more focus is to build sustainable operational frameworks to deploy our models.

The objective of this test is to assess your skills in these three aspects. The context is employment. Data profiles (scientists, architects, engineers) have a lot of similarities in their set of skills as the frontier between them is not clearly defined. Is a profile with skills in python and SQL a data scientist or data architect? Data candidates apply to many positions but often not to offers corresponding to their experience and skills. For HR teams, it can be a difficult task to identify or filter a profile based on skills because they are not aware of various technologies. Having a model that can help HR teams in their daily job would certainly be of interest.

Objective of the test

The objective of this exercise is to build and put in operation a machine learning model that can predict the profile of a candidate from useful pieces of information. From a dataset, you will:

1) Build a baseline ML model to predict the profile of individuals whose profession is not labelled – ML step
2) Deploy your ML model and demonstrate its operation – OPS step

The dataset for training and testing the model is provided in the file ***dataset_train_test.csv***. It contains a spreadsheet of ~10000 lines describing the profile of the candidates. This table is made up of 6 columns

- ***Entreprise*** which corresponds to a list of companies
- ***Metier*** which corresponds to the candidate's profession. This list contains the values: "Data scientist", "Lead data scientist", "Data engineer" and "Data architecte"
- ***Technologies*** which corresponds to the skills mastered by the profile
- ***Diplome*** which corresponds to their school degree (None, BSc, MSc, PhD, etc.)
- ***Experience*** which corresponds to the number of years of experience
- ***Ville*** which corresponds to the place of work

After deploying your model, you have 10 users who are eager to use your solution to predict profile of individuals.

Raoul would like to know the profile of:

- Thierry, PhD, 4 years of experience, living in Marseille and working for Symantec, who is skilled in Python, Tensorflow, scikit-learn, deep learning and R.
- Salomé, working for Jacobs since 2,5 years with proficiency in Python, Spark, GNU and Linux.
- Yannick from Ball Aerospace. He has a Master degree and 15 years of experience, and is experienced in VBA, Python, Excel and R.
- Jason, living in Toulouse and working for eHire LLC for 1,5 year after his Master. He is skilled in Java, C, C++, R and Python
- Kevin from Bordeaux, working with Map-Reduce/HDFS/PIG/HBASE/Python/Cassandra. He has some years of experience and holds a Master degree.

Ahmed would like to use the solution for a benchmark analysis on data profiles in Lyon and surroundings:

Helena	Working at Norfolk Southern Group in Grenoble	No experience	PhD	Skills on the cv: Python, Tensorflow, scikit-learn, Deep learning, R
Gabriel	KPMG in Lyon	2 years of experience	PhD	Python, Microsoft Azure, R, SQL appear as skills on the cv
John	Ashton Lane group in Lyon	5 years of experience	No diploma	Proficient in Python, Pyspark, Spark
Alain	Based in Lyon and working remotely for Google	3 years	Master degree	Python/R/machine learning/Excel/VBA/C++
Julien	Ball Aerospace in Lyon	1,5 year of experience	MSc degree	Working with technologies like Linux, Python, Hadoop, Perl, Ruby
Emilie	Starting a position at Turner in Saint-Etienne	6 months	PhD in physics	Excellent knowledge of Excel, Python, Matlab, R, machine learning. She speaks perfectly English

Bilal found three profiles with the same set of skills and would like to know which profile to assign to each:

Mireille, Leo and Teresa, all proficient in SAS, Teradata, SQL, R, Python, Machine learning and speaking English.

Mireille has been for Amazon for 1,5 year. She holds a PhD and lives in Toulouse

Leo celebrates his 8th year at Partners HealthCare(PHS), a tech company based in Bordeaux. He joined the company after his Master degree.

Teresa just completed her Master degree and started a position at KPMG in Marseille

Auguste would like to know if he made a good guess by saying that

Philippe, MSc, working for J.E. Ranta Associates in Rennes using Cassandra, MongoDB, NoSQL, AWS is a Data architect.

Sarah from Marseille, 4 years of experience working for Pearson and mostly doing R, Python, Spark, Pycharm, SAS, SQL, is a data scientist

Submission of results:

- ML step
  - You must use Python (ideally version 3)
  - You must use only the following libraries: pandas, numpy, matplotlib, sklearn, seaborn
  - You are expected to submit a Jupyter Notebook (.ipynb file) or your script (in a single .py file or a program) presenting all your work before the interview, which must be easily runnable by us.
- OPS step
  - You can use any framework in any language of your choice
  - You can use your local server or any cloud option
  - You are expected to showcase the pipeline you have designed and implemented. On our side the focus will be on model registry, model deployment, model monitoring.
- Presenting your work during the interview
  - You will present your work (context, objective, methodology, results, etc.) during your interview.
  - You have to prepare a presentation (PowerPoint or equivalent) of slides to fit in 25 minutes maximum.
  - You should prepare to present your work in front of an audience that is broad, combining experts and laypersons. It is also important to have a balanced presentation between the business opportunities and technical aspects of your solution.

Contacts:

Raoul Happy, AI Director

raoul@pivotandco.com