Machine Learning Engineer Challenge

Data Science team @ SulAmérica

Introduction

We have a vision - provide **integral health** by taking care of beneficiaries' **physical**, **mental** and **financial** health.

At the Data Science team, we create models that aim to facilitate our business stakeholders to reach this vision. But we won't deliver value if these models are not deployed in a robust machine learning system. We need to design, build, and keep improving this system.

That's where we hope you will fit in.

But first, you have to overcome this challenge.

And as you will surely find out, time'll be your worst problem

The Challenge

HL7 FHIR is an interoperability protocol for exchanging electronic health records. FHIR solutions are built from a set of modular components called "Resources". These resources can easily be assembled into working systems that solve real world clinical and administrative problems.

At SulAmérica, we use Google Cloud's Healthcare API as our FHIR solution. Real-time data is feeded to our FHIR server endpoint, registering Patient data, such as Observations, Encounters, and more.

You are given a sample of anonymized Patient records.

Your task is to design, build, generate and serve a diabetes identification model.

Given a user id, is he a diabetic?

Oh, and you've got one week to deliver.

A tad bit crazy, right? Yeah, maybe. But give it your best shot.

Focus on what you do best and just outline what would be the complete solution on the other parts.

We'll even help you. Here's a list of what is expected. So pick one or some topics and code away.

Of course, you are welcome to exceed our expectations!

- **Infrastructure**: system architecture, data ingestion, distributed processing, docker

- **Pipelines**: orchestration tools, data pipelines, feature engineering, model inference pipelines, CI/CD pipelines
- **Machine Learning:** model retraining, model selection
- **Evaluation**: choice of metrics, quality assurance, model monitoring
- **API**: interface, scalability, robustness, fallback

Inputs

We are providing you a zip file - https://storage.googleapis.com/workshop-python/output.zip **-** containing two folders, representing two views of the same data. `fhir` folder is the raw, json, FHIR data that you would find in Healthcare API, and `csv` folder is structured data that you would find in our data warehouse, powered by BigQuery.

You are welcome to design a system architecture from any of these two data sources. Our preference, as you can see, is Google Cloud, but you can design using resources from any cloud provider you are more familiar with.

Don't fret too much over using everything available, either row or column-wise. Occam's razor will be your friend in this challenge.

Documentation

We encourage you to write down every step down the road. Design decisions, references you used, difficulties and insights, and partial results.

That way, wrapping it all up into a presentation should be a breeze.

Results delivery

1. **All code you produce must be either on a jupyter notebook or in python files.**
   Organization is expected!
2. **You may use any language, tools, and cloud services you want.**
   But if you use Python we will like you more ;)
3. **After our evaluation, if you are selected, you will present your results in a video call.**