SPIKE

Desafío ML Engineer - Predicción de precios de insumos básicos en Chile

¡Gracias por participar en el proceso de selección de Spike! Como parte del proceso, este desafío nos ayudará a entender la manera en que te enfrentas a problemas y además podremos evaluar tus conocimientos.

Algunos puntos importantes a considerar:

- Este desafío no te debiera tomar más de 5 horas de tu tiempo (y ojalá menos!). Como es un desafío técnico que no tiene asociado una retribución, recomendamos no invertir más tiempo del recomendado.
- Tendrás hasta el **domingo 12 de diciembre a las 23:59 UTC-3** (hora chilena) para entregarlo.
- Habilitamos un canal en Gitter donde puedes conversar con el resto de las personas que están haciendo el desafío y con personas de Spike (por si tienes dudas). Ingresa en: https://gitter.im/ml-engineer-2021-2/desafioSpike (se requiere cuenta github o gitlab).
- Envía el desafío a través de este formulario: https://forms.gle/x9P5z9pFbYuScqiF6
- ¿Qué viene después? En un plazo máximo de una semana recibirás feedback. Y si pasas a la siguiente etapa, te invitaremos a conversar acerca del desafío, conocernos, contarte más de Spike y cómo sigue el proceso.
- Finalmente, ¡lee bien las instrucciones!

Introducción

Una de las tareas más frecuentes de un(a) ML Engineer es llevar a producción modelos de predicción, cuyo código fue creado inicialmente en una fase de exploración. El código comienza a crecer y poco a poco se empieza a desordenar. Finalmente, llega el momento de llevar el código a producción. Nos encantaría que el modelo estuviera en un servicio autoescalable, con una documentación tan fácil de entender que no tuviéramos que explicarle nada al cliente. Nos gustaría que consultar el modelo fuera tan fácil como hacer una request HTTP en un endpoint y el modelo nos entregara la respuesta. Además, nos gustaría tener logs y poder monitorear el modelo para entender sigue funcionando como esperamos a medida que pasa el tiempo.

Pero en realidad, lo único que tenemos son una serie de Jupyter Notebooks que usó la persona que construyó el modelo. Estamos contra el tiempo y necesitamos convertir esos notebooks en algo que cumpla lo mejor posible con nuestras

expectativas. Ojalá pudiéramos cumplir con todo lo que esperamos, pero debemos priorizar. Sería genial implementar algún tipo de monitoreo, pero levantar una API con el modelo es más urgente y el monitoreo puede esperar. Ahora bien, si alcanzamos a hacer todo en el tiempo que tenemos, mejor aún.

El desafío

En este desafío te vas a enfrentar con código real, escrito por un(a) Spiker cuando el/ella estaba postulando para entrar al cargo de Data Scientist. En esta ocasión el desafío consistía en construir un modelo de Machine Learning para predecir el precio de insumos básicos en Chile. Los/las postulantes nos debían entregar un Jupyter Notebook con el código de entrenamiento y predicción de este modelo. Ahora tu trabajo será empaquetar este código para que pueda ser puesto en producción fácilmente en cualquier plataforma (cloud u on premise).

En concreto, debes hacer lo siguiente:

1. Separar el código de entrenamiento del código de predicción.
2. Construir un pipeline de entrenamiento. Este pipeline debe tener al menos dos etapas claramente definidas. Por ejemplo: pre-procesamiento y entrenamiento. El resultado final de este pipeline debe ser un archivo con el modelo de ML serializado.
3. Construir una API que use el modelo previamente construido y que exponga un endpoint en el que se pueda consultar por las predicciones del modelo. Este servicio debe poder ser empaquetado en un contenedor y debes incluir todo el código para poder construirlo y ejecutarlo.

Herramientas

1. Pipelines

Puedes usar cualquier herramienta con la que te sientas cómodo(a), siempre y cuando esta sea open source y no dependa de un vendor particular. Por ejemplo, podrías usar Airflow, pero no Google Cloud Composer, porque esta última herramienta es privativa de Google Cloud.

Si bien no es un requisito mínimo, nosotros privilegiamos el uso de herramientas que puedan ser probadas fácilmente en un ambiente local y no necesiten de gran infraestructura (por ejemplo, un cluster) para ser ejecutados.

2. API

- Debes usar un framework que te permita crear una API para dejar el modelo en producción. El único requisito es que sea un framework open source.
- Debes incluir todo el código necesario para construir un contenedor que pueda ser ejecutado tanto en la nube, como localmente y dejar corriendo un servidor con esta API.

Entrega

Debes compartirnos un link a un repositorio git que contenga todo el código de tu desafío. Este repositorio deberá contener además un archivo [Readme.md] en el que expliques cómo abordaste el problema y cómo podemos ejecutar tu pipeline y tu API localmente.

Es muy importante hagas un buen Readme, para que alguien que no conoce tu código pueda ejecutarlo en su computador sin tener que hacerte ninguna pregunta extra.

Recuerda que la entrega es hasta **el domingo 12 de diciembre a las 23:59 UTC-3** (hora chilena) y será a través de este formulario:
https://forms.gle/16YQontt14jTKkoe9

Archivos base

En este repositorio encontrarás todo lo que necesitas para hacer el desafío:

- En este archivo encontrarás un Jupyter Notebook con la respuesta al desafío de alguien que actualmente es Data Scientist en Spike. En esta oportunidad, se les pidió a los postulantes, además de entrenar un modelo, construir visualizaciones y responder preguntas. Esto último no es relevante para tu desafío. Te puede servir para entender el contexto, pero no es necesario que lo incluyas en tu pipeline. Sí es necesario incluir cualquier tratamiento que se le haya hecho a los datos.
- En este archivo encontrarás las instrucciones del desafío de Data Scientist asociado a este notebook. Leerlo te servirá para entender el contexto.
- En la carpeta data encontraras los datos asociados al desafío.