PI Machine Learning Engineer Challenge

El problema

Nuestro equipo de data science ha estado desarrollado un modelo de regresión simple para predecir las ventas. Ya han logrado una versión aceptable de acuerdo a las métricas establecidas en conjunto con el negocio, por lo cual se quiere llevar a un ambiente de testing para su consumo. En necesario refactorizar el código que el equipo desarrolló. Para tal actividad se proporciona:

* Dataset con datos transaccionales con los cuales se entrenó el modelo.
* Jupyter notebook con el código en el cual se hizo el análisis y se entrenó el modelo.
* Json con un ejemplo de entrada al proceso de predicción.

Se necesita refactorizar y adaptar los siguientes pasos dentro del notebook a fin de poder crear un pipeline para procesar los nuevos datos de entrada:

* FEATURES ENGINEERING: para los años de establecimiento
* LIMPIEZA: Unificando etiquetas para 'Item\_Fat\_Content'
* LIMPIEZA: de faltantes en el peso de los productos
* LIMPIEZA: de faltantes en el tamaño de las tiendas
* FEATURES ENGINEERING: asignación de nueva categorías para 'Item\_Fat\_Content'
* FEATURES ENGINEERING: creando categorías para 'Item\_Type'
* FEATURES ENGINEERING: Codificando los niveles de precios de los productos
* FEATURES ENGINEERING: Codificación de variables ordinales
* FEATURES ENGINEERING: Codificación de variables nominales
* ENTRENAMIENTO

En caso de observar que algún procesamiento realizado por el equipo de data science no es reproducible o existe una mejor alternativa de solución, se puede implementar, hasta incluso volver a entrenar el modelo con tal objetivo.

Por otro lado, existe la posibilidad de que este proceso pueda ejecutarse de manera programada una vez a la semana, o también se está analizando la posibilidad de disponibilizar una api para su consumo. Esperamos que nos puedan mencionar algunas soluciones para implementar ambos escenarios, comentando que arquitectura se podría utilizar, así como las tecnologías y/o servicios que podrían estar involucrados y consideraciones a tener en cuenta.

Objetivo y presentación de los resultados.

* Se espera como entregable:
  + **Modelo guardado**: el modelo deberá serializarse, por ejemplo, como pickle(.pkl).
  + **Desarrollo del pipeline de procesamiento**: script/s que permita/n procesar los datos de entrada que están en formato json. Dentro de este pipeline se debería aplicar las transformaciones a los datos, cargar el modelo guardado anteriormente, y finalmente, retornar el valor de las predicciones.

Prueba: Debería mínimamente poder ejecutarse el script principal en consola, junto a la dirección de un json, y que se corra todo el proceso arrojando la predicción.

* + **Propuesta de arquitectura(opcional)**: descripción de la arquitectura y servicios involucrados para la implementación solución. Puede ser un boceto de las herramientas/servicios a utilizar, o un párrafo con la explicación de los componentes.

Queremos conocer detalles como los criterios utilizados, dificultades encontradas en el camino.

* Podes usar cualquier herramienta y/o servicio en la nube que desees.
* Nos gustaría ver tu código o implementación.
* Tendrás una presentación de 30 minutos para mostrar las transformaciones realizadas.
* Esperamos recibir tu trabajo en alrededor de 5 días posterior al envío del challenge.

Si tenés alguna duda, no dudes en contactarnos en [recursos.humanos@piconsulting.com.ar](mailto:recursos.humanos@piconsulting.com.ar)