Prueba técnica Data Science

En Los Ángeles existe un sistema compartido de bicicletas que brinda datos anónimos acerca del uso del servicio. La tabla que se proporciona contiene el histórico de viajes que se han realizado desde 2016 y contiene una columna que es de particular interés y que se buscará analizar a más profundidad: **Passholder_type**. A continuación se presentan las columnas que contiene la tabla:

- *trip_id:* identificador único para el viaje
- *duration:* duración del viaje en minutos
- *start_time:* dia/hora donde en viaje inicia en formato ISO 8601 tiempo local
- *end_time:* dia/hora donde el viaje termina en formato ISO 8601 tiempo local
- *start_station:* la estación donde el viaje inició
- *start_lat:* la latitud de la estación donde el viaje se originó
- *start_lon:* la longitud de la estación donde el viaje se originó
- *end_station:* la estación donde el viaje terminó
- *end_lat:* la latitud de la estación donde terminó el viaje
- *end_lon:* la longitud de la estación donde terminó el viaje
- *bike_id:* un entero único que identifica la bicicleta
- *plan_duration:* número de días que el usuario tendrá el paso. 0 significa un viaje único (Walk-up plan)
- *trip_route_category:* “Round trip” son viajes que empiezan y terminan en la misma estación
- *passholder_type:* El nombre del plan de passholder

Tareas a realizar:

1) Exploratorio de datos: Para comenzar la asignación se requiere realizar un análisis exploratorio de datos que busque contestar preguntas **relevantes** a los siguientes dos temas:
    - Saturación del servicio: La empresa busca contar con la disponibilidad más alta de servicio en el mercado, por lo que se quiere entender cómo se comporta la demanda de servicio en las distintas estaciones y horarios para cada plan
    - Crecimiento de planes: Se tiene la intuición que la tendencia en uso de bicicletas compartidas entre estaciones va a la alta, por lo que se requiere realizar una correcta planificación de bicicletas que deben tener. Adicionalmente, se espera que los planes de consumo anual crezcan en mayor proporción.

    En este punto se evalúa en storytelling y las conclusiones a que se llegan basado en la información proporcionada

2) Modelo analítico: Se desea saber si es posible inferir el tipo de pase tomando en cuenta las demás variables de viaje.
    - Construya un modelo analítico que incluya los puntos indispensables a considerar para un modelo (feature engineering, diseño de train-test split, cross-validation, métricas de desempeño, entre otros)
    - Interprete el resultado en contexto del problema y determine qué variables impactan en la predicción
    - Tomando en cuenta los exploratorios y el modelo analítico, ¿cree que es un buen modelo? ¿Qué variables adicionales añadiría para mejorar el modelo?

3) Evaluación del modelo

    Para la evaluación del modelo se realiza de manera automática a través de la plataforma de kaggle en la siguiente liga:

    https://www.kaggle.com/t/e82d8dd1223a4a459037106a2acab561

    En el tab de **Data** aparecen tres archivos:
    - *train_set.csv:* datos que servirán para construir el modelo
    - *test_set.csv:* datos que no tienen la variable dependiente y que se debe hacer predicción
    - *sample_submission.csv:* formato en el cual se debe subir las predicciones

    En el tab de **Submit Predictions** se hace upload de un archivo con el formato adecuado. **Favor de registrarse con algún pseudónimo que no esté relacionado con su nombre y compártenoslo.** El objetivo **no es llegar al mejor resultado en términos de la métrica de error**, si no estar seguros de tomar en cuenta aspectos indispensables al crear este modelo analítico. Por lo que sobrepasando el benchmark se considera una buena entrega.

4) Por último, se desea poner en producción el modelo. Agregue un diagrama del flujo completo incluyendo la puesta en producción. Si tiene experiencia con servicios de nube puede incluirlo.

Especificaciones técnicas requeridas:

- Incluir documentación del flujo de trabajo: A través de un documento en pdf explicar procedimiento correspondiente con cada tarea y subtarea. Es decir, agregar las visualizaciones resultantes del exploratorio de los datos, justificar la selección del o los modelos así como la interpretación de los mismos y demás puntos que consideres importantes para tu evaluación.
- Mantener en el código sólo lo necesario para la funcionalidad a realizar, quitar código comentado, así como librerías y código innecesario.

- Crear un repositorio en github, subir código y compartirnos tu url

Puntos extras:

- Utilizar docker para empaquetar código y dependencias.