Enunciado de Proyecto Final - Modelado Bicicletas Públicas GCBA

El gobierno de la ciudad de Buenos Aires (GCBA) desea modelar el comportamiento de los viajes de bicicletas públicas realizados utilizando datos hasta el mes de Agosto 2024 inclusive. El fin es poder predecir la cantidad de arribos de bicicletas que tendrá cada estación en un lapso de tiempo futuro conociendo la cantidad de partidas que tuvieron cada una de las estaciones en un lapso de tiempo pasado y considerando que no se conoce la estación destino de cada usuario cada vez que parte de una estación para iniciar un viaje. También, se asume que la cantidad de bicicletas es constante, es decir, que todas las bicicletas que parten de las estaciones siempre llegan a otras estaciones y siempre permanecen en el sistema. El impacto de aplicación de este caso radica en poder entender la logística de bicicletas de forma más inteligente con el fin de hacer mejor uso de los recursos del GCBA.

Conjunto de Datos

Dataset: https://data.buenosaires.gob.ar/dataset/bicicletas-publicas

- Recorridos Realizados 2020, 2021, 2022, 2023, 2024.
- Usuarios Ecobici 2020, 2021, 2022, 2023, 2024.

El dataset “Recorridos Realizados” se compone de la siguiente manera:

- Cada fila es un viaje realizado por un usuario y representa un viaje en bicicleta pública desde una estación origen a una estación destino.
- Cada usuario tiene un ID asociado y otras fuentes como género, modelo de bicicleta, fecha de alta al servicio entre otros.
- Cada viaje tiene un horario, día y mes de inicio en estación origen y un horario, día y mes de finalización en estación destino.
- Cada estación tiene factores como latitud y longitud, barrio, dirección entre otros.

A continuación se presentan el header y primeras 3 filas del archivo “*badata ecobici recorridos realizados 2024.csv*”.

id recorrido	duración recorrido	fecha origen recorrido	id estación origen	nombre estación origen	dirección estación origen
20428222	568	2024-01-23 18:36:00	513	308 - SAN MARTIN II	Av. San Martín 5129
20431744	1355	2024-01-23 22:41:20	460	133 - BEIRO Y SEGUROLA	Segurola 3194

20429936	0	2024-01-23 20:06:22	467	328 - SARMIENTO II	Sarmiento 2037

long estación origen	lat_estacion_origen	fecha destino recorrido	id estación destino	nombre estación destino
-58.490739	-34.5971297	2024-01-23 18:45:28	498	055 - HABANA
-58.51193	-34.6075	2024-01-23 23:03:55	382	204 - Biarritz
-58.3958925	-34.6055135	2024-01-23 20:06:22	6	006 - Parque Lezama

dirección estación destino	long estación destino	lat_estacion_destino	id_usuario	modelo bicicleta	género
Gral. José Gervasio Artigas 4298 (y Habana)	-58.4949585	-34.5865976	992557	FIT	MALE
Biarritz 2403	-58.477255218997	-34.6054308136996	320782	FIT	FEMALE
Avenida Martin Garcia, 295	-58.369758	-34.628526	828678	FIT	FEMALE

Tarea fundamental

Se conoce cuántas bicicletas salen de cada estación en cada ventana de tiempo. También tenemos registro de cuantas llegan a cada estación en cada momento. El objetivo de este caso es modelar el sistema de manera tal de predecir cuántas bicicletas van a llegar a cada estación en los próximos ΔT mins a partir de un mes-dia-hora dado teniendo como input cuantas bicicletas salieron en los últimos ΔT mins (puede ser un ΔT a definir por el desarrollador del modelo).

De forma más específica: Dado un punto del tiempo T (queda a criterio del usuario) , observando el ultimo ΔT pasado (por ejemplo: T - 30 m) de cantidad de partidas de bicis de cada estación + perfil del usuario + ubicación de la estación + las features que considere conveniente -> predecir la cantidad de N arribos que tendrá cada estación en el próximo ΔT futuro.

Por ejemplo: Dada la cantidad de partidas de bicicletas conocidas en los últimos 30 minutos en todas las estaciones de BA Bicis: predecir la cantidad total de arribos que tendrá cada estación en el lapso de los próximos 30 minutos. Una posible forma de estructurar los datos

Variable aleatoria Input X:

- Mes, dia, rango horario de la ventana de tiempo
- Barrio estación A
- Latitud y longitud estacion A
- Cantidad de despachos/partidas de la estación A en los últimos 30 mins
- Detalle del usuario de cada partida de la estación A
- …
- Barrio estacion Z
- Latitud y longitud estacion Z
- Cantidad de despachos/partidas de la estación Z en los últimos 30 mins

- Detalle del usuario de cada partida de la estación Z

Variable target Y:

- Cantidad de usuarios que van a llegar a la estación A en los próximos 30 mins
- …
- Cantidad de usuarios que van a llegar la estación Z en los próximos 30 mins

Consideraciones

1. Los datos utilizados para train, validation y test mandatoriamente deben utilizarse **HASTA** Agosto 2024 inclusive. Es decir que para todo el análisis solo se pueden usar cualquiera de los datos de 2020, 2021, 2022, 2023 y hasta Agosto 2024 inclusive.
2. El dataset como esta requiere de preprocesamiento para construir los samples X y las labels Y. Queda a criterio de los desarrolladores las decisiones de preprocesamiento para obtener las samplers X y las labels Y.
3. Respecto a la temporalidad de los datos:
    a. Se puede armar un dataset donde cada muestra X es un vector que contempla la cantidad de despachos que tuvo cada estación y sus respectivas features en la ventana ΔT histórica de tiempo determinada.
    b. Por otro lado se puede armar un dataset donde cada etiqueta Y es un vector que contempla la cantidad de arribos que tuvo cada estación en una ventana ΔT futura de tiempo determinada.
    c. Entonces la idea es aprender un modelo f que mapee comportamiento de despacho de bicicletas durante una ventana de tiempo pasada hacía comportamiento de arribo de bicicletas durante una ventana de tiempo futura para todas las estaciones.

Ideas para extensiones

1. Hacer un mapa de calor para distintos rangos de tiempo para ver la intensidad de los viajes de bicicletas entre distintos pares de estaciones.
2. Probar distintos valores de ΔT histórico para usar como input de predicción y probar otros valores distintos de ΔT futuros para predecir. Es decir, el ΔT histórico que entra al modelo no tiene que necesariamente ser igual al ΔT de valores futuros.
3. Revisar que el GCBA tiene mas datasets separados con features de los usuarios y estaciones en https://data.buenosaires.gob.ar/dataset/bicicletas-publicas. Ver de concatenar los datasets que considere necesarios para enriquecer de features al modelo.
4. ¿Hay alguna feature que sea más importante que otras para predecir?
5. ¿Hay alguna estación que sea más fácil de predecir que otras?
6. ¿Es mejor armar 1 solo modelo que integre todas las predicciones o armar múltiples modelos por separado?