Máster en Tratamiento Estadístico y Computacional de la Información

Minería de Datos

2019/2020

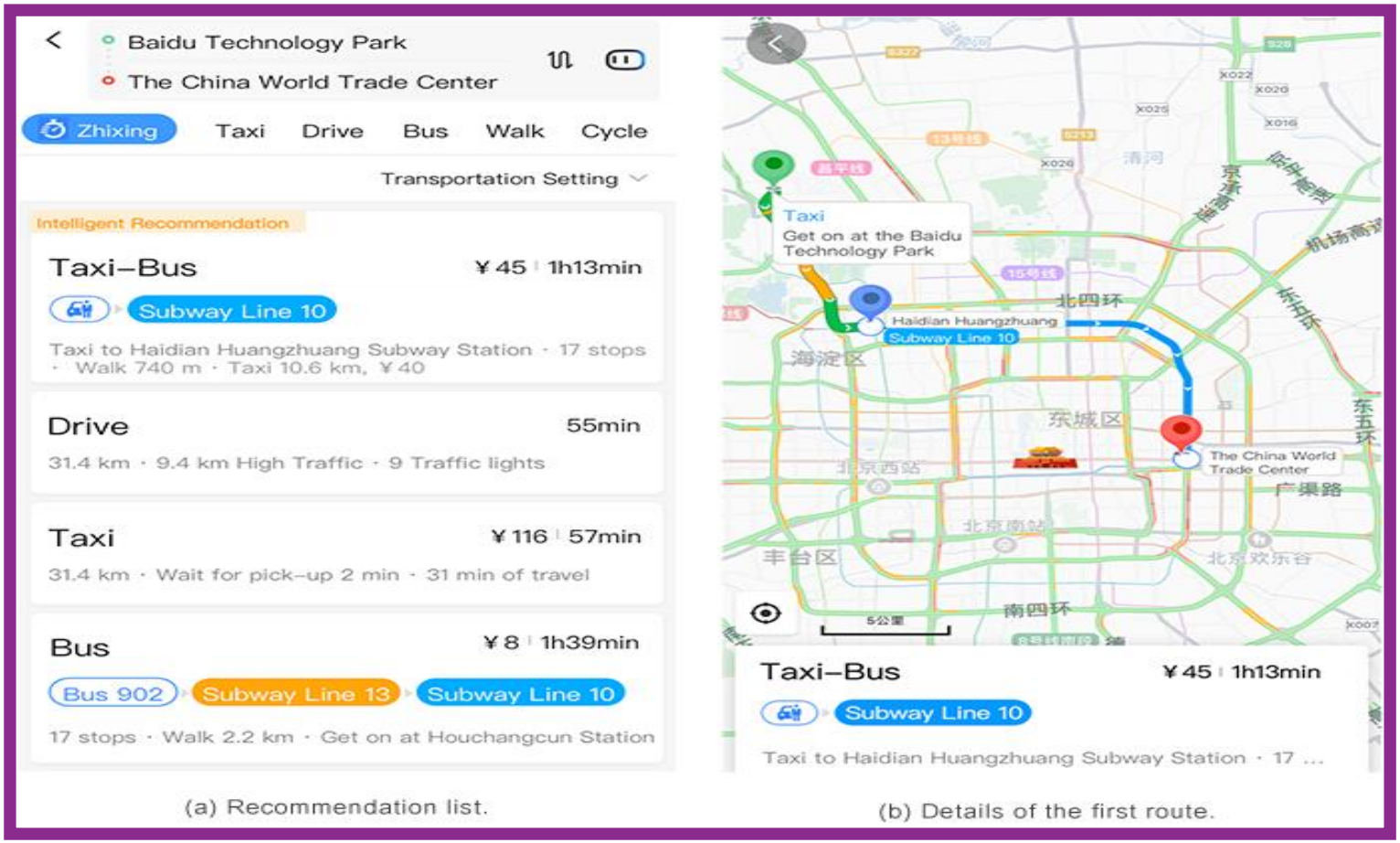
Sistema de recomendación de transportes

Daniel Vélez Serrano

DVS

Sistema de recomendación de transportes

- El objetivo de la práctica es **predecir la opción elegida por un usuario al que se le presentan diferentes posibilidades de transporte** para moverse entre dos puntos de una ciudad china.



Sistema de recomendación de transportes

- La **tabla** que se proporciona **consta de las siguientes variables:**

  - SID: Identificador de sesión abierta por el usuario (variable ID).
  - Date: Fecha en la que realiza la búsqueda.
  - Hour: Hora a la que realiza la búsqueda.
  - X_origin_0.- longitud del punto de origen.
  - Y_origin_0.- latitud del punto de origen.
  - termicSensation_origin.- sensación térmica en el punto de origen (ºFarenheit).
  - X_destination_0.- longitud del punto de destino.
  - Y_destination_0.- latitud del punto de destino.
  - termicSensation_destination.- sensación térmica en el punto de destino (ºFarenheit).
  - transport_mode.- medio de transporte ofrecido al usuario en la sesión.
  - distance.- distancia asociada al medio de transporte transport_mode.
  - price.- precio asociado al medio de transporte transport_mode. Puede ser 0 (andando, bici, etc.).
  - eta.- duración (tiempo) asociado al medio de transporte transport_mode.
  - min_distance.- mínima distancia ofrecida al usuario en la sesión.
  - min_price.- mínimo precio ofrecido al usuario en la sesión. Puede ser 0 (andando, bici, etc.).
  - min_eta.- mínima duración (tiempo) ofrecida al usuario en la sesión.
  - binaryTarget.- variable a predecir identificadora de si el usuario hizo click (1) o no (0) en la opción ofrecida en la sesión para cubrir el recorrido de (X_origin_0, Y_origin_0) a (X_destination_0, Y_destination_0) (variable target)..

Sistema de recomendación de transportes

- Se pide **ajustar modelos para predecir si un usuario hará *click* en alguno de los medios de transporte que se le presentan.** Los **tipos de modelos** a contrastar serán **árboles de decisión, regresiones logísticas, redes neuronales y ensamblados** (*random forest* y *gradient boosting*). La tabla a utilizar para ajustar los modelos es: *tablatransportestrain.sas7bdat*.

- Se deberá **entregar:**

  - **Un documento WORD/PDF** en el que, para **cada tipología de modelo**, se muestre **aquél que proporcione mejores resultados**, **DETALLANDO** las **opciones especificadas** para el mismo (posibles sobre/bajomuestreos, tratamiento de *missings/outliers*, opciones del modelo en cuestión, etc.), así como una **interpretación de sus resultados** (reglas de los árboles, estimaciones de los parámetros, gráficos de iteración en redes y modelos ensamblados, etc.).

  - Un archivo (SAS/CSV) con la asignación que se hace (0/1) a cada uno de los SIDs de la tabla *tablatransportes_eval_inputs.sas7bdat.* Dicho archivo debe contener únicamente dos columnas: el identificador de SID y la clase asignada (0/1).

- El informe se puntúa sobre 5. Además, cada equipo recibirá:

  puntuación adicional = (número de equipos - posición equipo+1)/(número de equipos)

- La **fecha límite** de entrega de la práctica es el **2 de marzo a las 23:59:59.**