Predicción de Flujos en un sistema de Bikesharing (Ecobici)

Denisse Martínez & José A. Ramírez
ITAM

26 de mayo de 2016

Los sistemas de bikesharing se han expandido en los últimos años en las ciudades altamente congestionadas y con ello se han generado nuevos retos para ofrecer servicios eficientes. En el siguiente documento se describe un modelo Probit para predecir el flujo en las estaciones de Ecobici de la Ciudad de México.

1. Antecedentes

Ecobici es un sistema de bicicletas públicas de cuarta generación, que implementó el Gobierno del Distrito Federal como parte de la Estrategia de Movilidad en Bicicleta. Desde la puesta en marcha el 16 de febrero del 2010, es gestionado por la Secretara del Medio Ambiente del Distrito Federal.

Inició operaciones a través de 85 cicloestaciones (1,114 bicicletas) separadas por una distancia de 300 metros entre una y otra en la cual daba servicio a las colonias Cuauhtémoc, Juárez, Roma Norte, Hipódromo Condesa y Condesa. En octubre 2011 se ampliaron operaciones al Centro Histórico de la Ciudad de México con 5 cicloestaciones y 12 reubicaciones de las existentes, además de Polanco y zonas aledaas. En 2015 contaba con 444 cicloestaciones (6,262 bicicletas) con un área de cobertura de 32 km2 en 42 colonias de las Delegaciones Benito Juárez, Cuauhtémoc y Miguel Hidalgo.

Ecobici, una alternativa de movilidad, funciona como eficaz complemento a los sistemas de transporte y ayuda a resolver problemas de movilidad en la Ciudad de México que es una de las ciudades más grandes del mundo.

El sistema es operado por Clear Channel Outdoors a travs de su división SmartBike, quien alrededor del mundo ha implementado sistemas similares en España, Francia, Noruega, Suecia e Italia.

1.1. Problema a resolver

Si los gobiernos desean impulsar la movilidad en bicicleta (Ecobici) y generar una alternativa real de transporte, resulta fundamental ofrecer un servicio eficiente, en donde el usuario pueda tener la certeza que encontrará biciletas disponibles cuando asista a una estación y así mismo, espacios disponibles cuando desee entregar una bicicleta. Este estudio busca crear una solución que permita a los operadores anticiparse a la demanda de bicicletas y espacios disponibles para los arribos. El siguiente análisis es una primera fase de un proyecto más amplio que busca incorporar información en tiempo real.

Este estudio tiene como predecente un análisis de comunidades en el que se analizó que tan conectada está la red y cuáles son las áreas de flujo que conforman las regiones operativas del sistema Ecobici. Se analizaron distintos horarios y se observó que durante los días lunes a viernes y entre las seis y doce del día, la red se agrupa en trece regiones. Estos resultados indican una alta concentración a comparación de resultados observados para los fines de semana, en donde la red se divide hasta en veinticinco regiones operativas. Las trece regiones entre semana en el

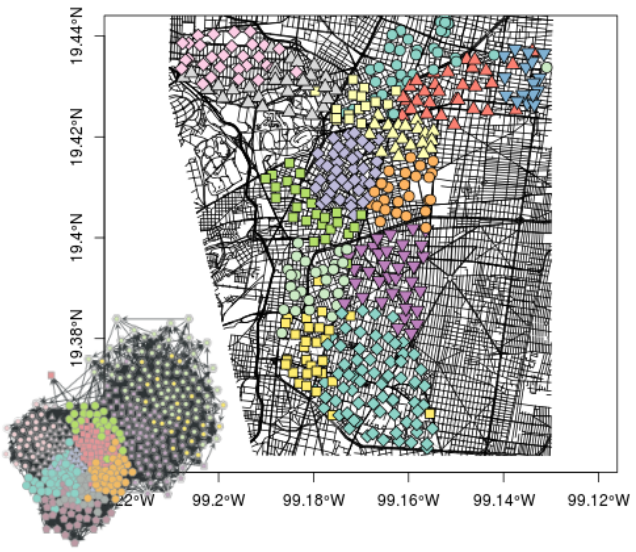

Figura 1: *Distribución de las regiones de Ecobici*

horario en la mañana agrupan zonas como Polanco - Oeste, Polanco - Este, Cuauhtémoc, San Rafael - Tabacalera, Juárez, Centro, Roma Norte, Roma Sur, Condesa, San Miguel Chapultepec - Escandón, Nápoles - San Pedro de Los Pinos, Insurgentes - Mixcoac, Narvarte - Del Valle Norte, Del Valle Sur y Centro. Para mayor detalles del estudio, revisar documento anexo Análisis_comunidades.

1.2. Objetivo del Proyecto

Estimar un modelo Probit para cada región con el fin de predecir si existe un balance (Y = 0) o desbalance (Y = 1) en el stock de bicicletas de una estación. El alcance del proyecto se limita a las predicciones para los días entre semana en un horario de seis a doce del día. El análisis por región se realiza contemplando que cada una tiene características partículares latentes que pueden arrojar información relevante al modelo de prediccíon de cada una.

2. Metodología

El proyecto se puede consultar en el repositorio de GitHub: DennyMtz/DPA_Ecobici

- Análisis exploratorio de datos disponibles

  Las fuentes de información del proyecto son los datos públicos de Ecobici que dan información a nivel viaje sobre la estación de retiro, estación de arribo, sexo del usuario, edad del usuario, duración del viaje, con detalle en la fecha y hora. Otra fuente de información es la Dirección de Monitoreo Atmosférico de la Ciudad de México. Esta institución tiene distintas estaciones de monitoreo en el Valle de México mediante las cuales captura información de monóxido de carbono (CO), óxido nítrico (NO), dióxido de nitrógeno (NO2), óxidos de nitrógeno (NOX), ozono (O3), partículas menores a 10 micrómetros (PM10), partículas menores a 2.5 micrómetros (PM2.5) y dióxido de azufre (SO2). Asi mismo, se coleccionaron datos de humedad relativa, temperatura ambiental, velocidad y dirección del viento.

  Los datos de Ecobici fueron descargados de la página https://www.ecobici.df.gob.mx/es/estadisticas y los datos meteorológicos y de contaminantes de http://www.aire.df.gob.mx/. Los datos agrupan información de 2015 de 444 estaciones de Ecobici.

- Limpieza, creación y selección de atributos

  Los datos sobre viajes fueron agrupados por día considerando solamente los realizados entre semana, horas de seis a doce del día. Los datos de contaminantes y meteorológicos se aguparon por el nivel máximo de cada medición. Para específar más el modelo se agregaron variables binarias para las horas 6, 7, 8, 9, 10 y 11 y para los días entre semana lunes, martes, miércoles, jueves y viernes.

  Se calculó la correlación entre las covariables y se encontró que las variables de distancia promedio y duración promedio están correlacionadas en un 82 %, mientras que la edad promedio está correlacionada con esas mismas variables con un 63 % y 57 %, respectivamente. Dado a la naturaleza predictiva de este modelo, se omitirán dichas relaciones.

  La variable a predecir se construyó a través de estimar cada diez minutos el flujo neto: (bicicletas arribo - bicicletas retiro), obteniendo un promedio por hora ponderando por la capacidad máxima de bicicletas por estación. Posteriormente esta variable es considerada con valor "0"si su valor se encuentra entre (-0.1 y 0.1), es de-

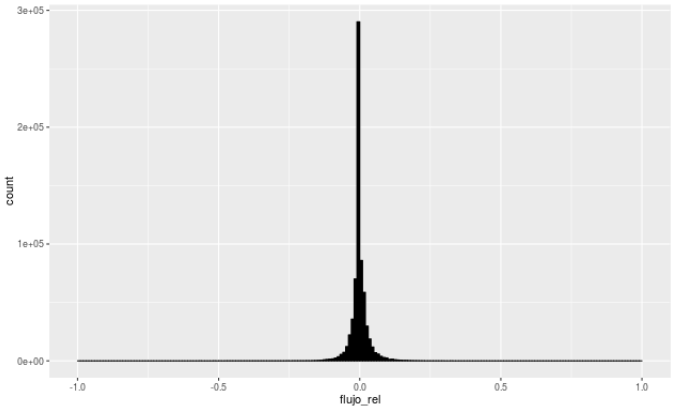

Figura 2: *Distribución del flujo relativo*

cir que está en balance y el resto toma el valor de "1", que implica un desbalance en cuanto al stock de bicicletas por estación, ver Figura 2. La variable de flujo se construyó de dicha forma, ya que no se cuenta con el dato histórico, sin embargo es un dato que se puede capturar en tiempo real a través de una API, información que será empleada en una segunda fase de este análisis.

- Definición del Modelo Probit

  El modelo Probit es un modelo en el cual la variable a explicar solamente puede tomar valores 0 y 1, estimando la probabilidad de pertenecer a una categoría u otra. El modelo representa una distribuciíon de probabilidad acumulada para una distribución normal.

  El modelo Probit desde una perspectiva de variable latente:

$$\begin{aligned} P &= (Y_1 = 1|x) = P(y_i^* > 0|x) \\ &= P(x_i^t\beta + \varepsilon > 0|x) \\ &= P(\varepsilon_i > -x_i^t\beta|x) \\ &= 1 - F(-x_i^t\beta) \end{aligned}$$

  Se asume que los errores son independientes y se distribuyen de forma normal.

  Cabe destacar que la estimación del modelo es bajo el contexto de análisis bayesiano, por lo que se estiman distribuciones a priori de los coeficientes del modelo. Dichas distribuciones se generaron a partir del Estimador de Máxima Verosimilitud y se emplea el método de Gibbs Sampling para obtener las distribuciones posteriores.

Se estimaron trece modelos Probit para las regiones definidas, cada uno especficado con 21 variables:

- Categorización de usuarios: edad promedio y proporción de usuarias de sexo femenino.
- Dinámica de los viajes y estación: duración promedio de los viajes, distancia promedio de los viajes y capacidad máxima de las estaciones.
- Contaminantes: Máximo valor registrado de O3 y PM10.
- Meteorológicos: temperatura máxima registrada durante el día, humedad relativa (RH), velocidad del viento (WSP) y dirección del viento (WDR).
- Binarias para las días de la semana: lunes, martes, miércoles, jueves y viernes La base de la estimación es el día lunes.
- Binarias para las horas de la mañana: 6, 7, 8, 9, 10 y 11.

3. Resultados

Los resultados muestran en promedio un 70 % de aciertos en la predicción de cada modelo por región. Para cada modelo se realizaron 1,000 iteraciones de Gibbs Sampler, obteniendo los coeficientes finales, de los cuales se deriva el análisis de resultados. Considerando la variable binaria lunes como base del modelo, los hallazgos más relevantes:

- La edad promedio, la proporción de usuarios de sexo femenino y la distancia promedio, aportan positivamente a la probabilidad de desvalance a lo largo de todas las regiones.
- En 10 de 13 regiones, la capacidad máxima contribuye a disminuir la probabilidad de desvalance.
- Los días martes, miércoles y jueves contribuyen positivamente a la probabilidad de desvalance en prácticamente todas las regiones.
- En las regiones 9, 10, 11 y 12 se observa un signo negativo en la probabilidad de desvalance para todas las horas de la mañana.

- Se muestra una probabilidad positiva en las horas 7, 8 y 9 en las regiones 0 a 8. Esto es esperado dado que a dichas horas de la mañana se realizan los traslados a los lugares de trabajo.
- En 9 de las 13 regiones, se muestra que la humedad relativa tiene un impacto negativo en la probabilidad de desvalance.
- En general, las variables de velocidad y dirección del viento contribuyen positivamente a la probabilidad de desvalance.
- En cuanto a las variables de contaminantes O3 y PM10, no muestra un patrón homogeneo entre las regiones.

Los resultados se muestran a detalle en los documentos adjuntos en los que se incluye una carpeta de cada región que contiene las distribuciones de coeficientes obtenidos mediante Gibbs Sampler.

Referencias

[Albert and Chib, 1993] Albert, James H. and Chib, Siddhartha (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*.