Predicción del tiempo esperado de recorrido en la Línea 1 del Metrobús de la CDMX

Raquel Yunoen Badillo Salas y Juan Javier Santos Ochoa

Resumen—El objetivo de este trabajo es predecir los tiempo de recorrido en el Metrobús. Ocupamos un modelo de árboles de decisión de XGBoost con una función Poisson. Ocupamos variables de interés como predictores (distancia, estación, tiempo del último metrobus, tráfico, número de estaciones por recorrer y si está lloviendo). Los resultados son mostrados una aplicación interactiva: https://goo.gl/RjR53s.

I. INTRODUCCIÓN

A Ciudad de México es la quinta ciudad con mayor población en mundo con 21 millones 581,000 habitantes y debido a esto uno de sus mayores problemas es la congestión en los sistemas de transporte. Esto acompañado con poca información de los horarios y tiempos aproximados de espera agravan el problema de congestión.

Uno de los sistemas de transporte más organizados en la Ciudad de México es el Metrobús, sin embargo, no existe ninguna fuente oficial de horarios de Metrobús que ayude al usuario a decidir su ruta. El objetivo de nuestro proyecto es ayudar a los usuarios a saber cuánto tardarán en el Metrobús y en cada estación. Esto contribuirá a que tomen mejores decisiones de transporte. De igual forma, las autoridades encargadas del sistema Metrobús también pueden contar con información oportuna para planear las operaciones.

En este proyecto creamos un modelo estadístico que calcula el tiempo esperado que demorará un viaje entre diferentes estaciones de la Línea 1 de Metrobús. El modelo usa datos históricos y en tiempo real para hacer estimaciones más precisas. Los resultados pueden ser consultados en este enlace: https://goo.gl/RjR53s.

El enfoque que aquí desarrollamos usa datos abiertos y puede ser fácilmente extendido a otras líneas del sistema e incluso se puede usar para la planeación de nuevas rutas alternativas ya que permite hacer extrapolaciones a cualquier punto geográfico de la ciudad.

II. DATOS

En este documento ocupamos los datos abiertos en la página de Hack CDMX para la linea 1 del Metrobús y proxies para mejorar la estimación. Estas bases contienen datos del identificador del viaje, localización, hora, fecha, identificador del vehículo, velocidad y horario de salida.

La base de datos que utilizamos fue la liberada el día 15 de noviembre del presente año en la página de Hack CDMX, que contiene la información de la operación de los buses de la línea 1 del día 14 de noviembre en el horario de las 16:20 hasta las 23:59. Inicialmente la base contaba con aproximadamente 70 mil observaciones, sin embargo, restringimos las observaciones a aquellas que contaban con un identificador del vehículo y del viaje, ya que esto era fundamental para nuestro enfoque. También eliminamos las observaciones que estaban fuera del área geográfica de influencia y las que tenían comportamientos inusuales, como por ejemplo, que con muy pocos segundos de diferencia estaban en lugares muy distantes. Tras el proceso de depuración nos quedamos con casi 35 mil observaciones válidas para el análisis.

1

Nuestro enfoque consistió en analizar el desplazamiento de los buses a nivel geográfico y en el tiempo para un mismo viaje. Un viaje es un trayecto de ida y regreso de un Metrobús desde su base inicial. La cantidad de observaciones que estaban disponibles para cada viaje eran variables, así como la posición y el tiempo en el que estas medidas se recolectaban. Si representamos cada observación de la siguiente forma, donde:

$O_{v,t,p}$

Es la observación para el viaje v, en el momento t, en la posición p. Nuestro análisis consistió principalmente en comparar las observaciones que pertenecían a un mismo viaje con respecto a la observación del periodo anterior $(O_{v,t_1,p_1}$ vs $O_{v,t_2,p_2})$. De esta forma calculábamos cuánto varió el tiempo (t_2-t_1) y la posición $(dist(p_1,p_2))$ entre dos observaciones consecutivas. Con este procedimiento pudimos obtener dos variables importantes para nuestro modelo: tiempo del recorrido y distancia recorrida.

Utilizamos la información disponible en la base de datos y la combinamos con otras fuentes para crear variables proxies que nos ayudaron a especificar el modelo. Estas variables explicativas son: 1) distancia Calculamos la distancia lineal entre los dos puntos de las estaciones de Metrobús. 2) Intersecciones El número de intersecciones viales, es decir, cuantos cruces debía pasar el vehículo de estación a estación. 3) Sentido Hacia qué sentido va el viaje (norte-sur o surnorte). 4) Velocidad promedio de los últimos viajes Es la velocidad promedio de los metrobuses que cubrieron el mismo travecto durante la última media hora 5) Tráfico Con ayuda de la API de Google Maps recolectamos datos de cuanto tráfico hay usualmente durante la hora del recorrido. 6) Número de estaciones el número de estaciones iniciales y finales en el recorrido y finalmente, 8) lluvia creamos una variable que indica el nivel de precipitaciones en la Ciudad de México a la hora del recorrido con datos de metroblue ¹

¹https://www.meteoblue.com

III. MODELO

Para predecir cuántos segundos tardaría un viaje desde un punto de inicio a un final, ocupamos árboles de regresión Boosting con una distribución Poisson. Los árboles de regresión son modelos estadísticos no lineales que buscan los mejores predictores correlacionales y no causales para una variable.

La variable dependiente de nuestro modelo es el tiempo que dura un recorrido y para entrenar el modelo, ocupamos como predictores las variables obtenidas en la sección anterior con sus interacciones obteniendo un total de 9843 variables con 34667 datos para entrenar el modelo. Obtuvimos un error cuadrático medio (MSE) de 0.6745. Es decir, nuestras estimaciones tienen un promedio de 0.6745 segundos de error.

Para todo el proceso de recolección, depuración, procesamiento de datos, estimación del modelo y creación de la aplicación web usamos software libre.

IV. APLICACIÓN

Los resultados de nuestro modelo los utilizamos para construir una aplicación web que nos permite consultar tiempos de traslado entre dos estaciones. Para ocupar la aplicación el usuario deberá seleccionar tres opciones: hora de salida, estación de origen y estación de destino. Con base en estos datos nuestra aplicación puede predecir cuanto tiempo se tardará en llegar al destino final, desagregando el resultado por cada estación que hace parte del recorrido.

El resultado de la aplicación final la podemos observar en la figura 1. La figura es interactiva entonces el usuario fácilmente puede saber cuánto tardará entre cada estación.

Un aspecto a destacar es que las predicciones tienen en cuenta información en "tiempo real", lo que permite tener estimaciones más precisas que si solo usáramos datos históricos. Esta información incluye el estado actual del tráfico, el nivel de lluvias y los recorridos.

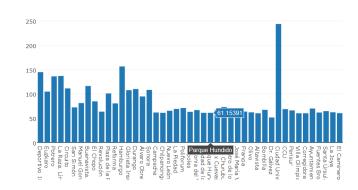
De igual forma, las predicciones son dinámicas ya que tienen en cuenta los tiempos estimados previamente para calcular el tiempo que tomará llegar a las siguientes estaciones. Por ejemplo, si un usuario calcula el tiempo esperado para la ruta "Indios Verdes - El caminero" el aplicativo primero calculará cuanto demorará en ir de Indios Verdes a la siguiente estación, Deportivo 18 de marzo, luego para estimar la duración del siguiente trayecto, de Deportivo 18 de marzo a Euzkaro, tendrá en cuenta el tiempo que tomó llegar a esta estación y actualiza la información de tráfico, lluvia y viajes recientes para la hora a la que espera que llegue a esa estación. Creemos que este enfoque es mucho más realista ya que las condiciones de movilidad de la ciudad pueden cambiar drásticamente a medida que el viaje se va desarrollando.

Debido a la disponibilidad de datos, el modelo fue construido sólo con información correspondientes al día miércoles 14 de noviembre, en horario de 16:20 hasta las 23:59. Por ello, limitamos las consultas a este horario y advertimos que probablemente las estimaciones son más adecuadas para predecir los tiempos de un miércoles en este rango horario. Si contáramos con información para todos los días de la semana en todo el horario de funcionamiento del sistema el modelo sería fácilmente generalizable.

Figura 1. Aplicación tiempo de espera Metrobús-Línea 1



El tiempo esperado de su viaje de Indios Verdes a El Caminero es de 64.51 minutos. A continuación se muestra el tiempo estimado (en segundos) entre estaciones:



V. Conclusión

En conclusión podemos mejorar la información que tiene el usuario del tiempo estimado de recorrido. La ventaja de nuestra aplicación frente a otras aplicaciones es que el usuario fácilmente puede acceder a distintos horarios con información de tiempo real para saber el tiempo de su trayecto y así optimizar mejor su viaje. Este modelo puede ser fácilmente extendido a otras líneas de metrobús y servicios de transporte público.