Propuesta de Investigación EEA-UBA

Uso de árboles de regresión para predecir el caudal de pasajeros en subterráneos de la Ciudad de Buenos Aires

Tokman Ezequiel; Saint Germain José; Clas Giulia

1. Resumen

En este trabajo se abordará la problemática del aglomeramiento de pasajeros en el sistema de subterráneos de la Ciudad de Buenos Aires. Para este fin, se generará un modelo basado en árboles de regresión que permita estimar el caudal de pasajeros de cada estación en un momento dado (septiembre) utilizando el algoritmo rpart (lenguaje R) y datos provenientes de la base de datos abierta del Gobierno de la Ciudad de Buenos Aires.

2. Problema

Una de las mayores problemáticas del uso de subterráneos como medio de transporte en la Ciudad de Buenos Aires está asociada al aglomeramiento de pasajeros. Esto trae aparejadas consecuencias tanto en el bienestar de los usuarios como en los requerimientos elevados de tarea de limpieza y mantenimiento de estos sectores. Debido a esta problemática, nos propusimos como objetivo general de la investigación obtener un modelo que permita predecir el volumen de pasajeros de cada estación de la línea de subterráneos de la Ciudad de Buenos Aires en un determinado momento del día. Obtener un modelo de estas características permitirá no solo adecuar la inversión en infraestructura de manera adecuada sino también planificar posibles ampliaciones de la red. Como se va a intentar predecir una variable numérica (cantidad de pasos por los molinetes de las estaciones), hablamos de un problema de regresión.

Para abordar esta investigación se utilizará la base de datos de la Ciudad de Buenos Aires, en donde se muestra el uso de cada molinete en cada estación de la red. También se adicionarán otras fuentes que permitan generar nuevos atributos al dataset que aumenten la performance del modelo. Como modelo predictivo, planteamos la utilización del algoritmo rpart disponible en lenguaje R, así como también generaremos un modelo base con las funciones integradas de R para generar regresiones lineales múltiples.

3. Objetivos

Como objetivo particular se busca que la persona que vea los videos que generaremos a partir de esta investigación adquiera una comprensión inicial de cómo plantear un problema de regresión aplicado en el área del transporte público. También se busca encontrar una aplicación práctica de un algoritmo de *machine learning* en el área de las políticas públicas. Con respecto a los árboles de decisión, buscaremos que la persona que vea los videos profundice en la manera en que este algoritmo genera predicciones, así como en los riesgos o limitaciones que pueda tener respecto al sobreajuste o a su poder de predicción frente a modelos más sofisticados.

4. Estado del Arte

Entre los trabajos previos a destacar podemos mencionar el trabajo de Cardozo, García Palomares y Gutiérrez Puebla (2008) en donde se busca estimar la demanda del metro de Madrid

con un modelo de regresión lineal múltiple, tomando como variables tanto fuentes externas (económicas, demográficas, entre otras) como (como el precio del boleto).

Beramendi, Gianotti y Vanesa (2019) se aproximaron al estudio de la red de subtes de Buenos Aires mediante un método cualitativo, que son los grupos de discusión o focus group. En los mismos, los participantes destacaron la rapidez, el precio de los boletos y la facilidad para llegar al centro de la Ciudad (la cual está bloqueada para vehículos particulares la mayor parte del día hábil), mientras que destacaron como molestia el tiempo de espera, la imprevisibilidad del servicio y la aglomeración de gente durante el transcurso de su viaje.

Por último, el artículo de Alkherebi et al (2023) implementa diversos modelos de aprendizaje automático para predecir el número de pasajeros de un sistema de metro; en este caso en el del estado de Qatar. En el mismo implementa y compara una diversidad de algoritmos (*Regresión Ridge*, *Support Vector Machine*, *Random Tree*, *KNN*, *Gradient Boosting Machine*, entre otros). En este paper se destaca la performance del árbol azaroso, el cual obtuvo un R cuadrado de 87.4%.

El texto de Akherebi nos da un punto de inicio para tomar los métodos utilizados en Qatar para el sistema de subte de la ciudad de Buenos Aires, mientras que las ideas obtenidas por Beramendi et al nos brindará ideas para generar atributos en nuestro dataset. No hemos encontrado trabajos previos que hayan utilizado específicamente el uso de molinetes en el sistema de subtes, por lo que este trabajo tiene la potencialidad de aportar un enfoque cuantitativo y exhaustivo de la manera en qué se comporta la demanda de la población del Área Metropolitana de Buenos Aires.

5. Metodología

5.1 Dataset: fuente, preprocesamiento y estructura

El dataset proviene de la base de datos abierta de la Ciudad de Buenos Aires. De la misma se obtiene una serie de datasets en formato csv con la cantidad de pasajeros por molinete cada quince minutos en una hora determinada en cada estación. Cada dataset tienen la siguiente estructura de atributos:

Nombre	Tipo	Descripción
fecha	fecha	Hora de registro
desde	tiempo	Hora de inicio del registro
hasta	tiempo	Hora de finalización del registro
línea	texto	Línea de subte
molinete	tiempo	Línea de subte con la identificación del molinete
estación	tiempo	Estación del subte
pax_pago	número entero	Pasajeros que pagaron

pax_pases_pagos	número entero	Pasajeros que pagaron con pases
pax_franq	número entero	Pasajeros que pagaron con franquicias
pax_total	número entero	Cantidad total de pasajeros que usaron el molinete

Para trabajar con este dataset vamos a discretizar la variable de hora, dividiéndola entre hora pico (momento de alta utilización del sistema) y valle (el resto de horas). Queda por definir si definiremos el mismo rango de horas para todas las líneas o si estimaremos la hora pico y valle de cada línea observando su distribución. Vamos a evaluar el modelo en el mes de septiembre del año 2023 entrenando al algoritmo con los datos de septiembre de 2022, 2021, 2020, 2019, 2018 y 2017.

Con respecto a la transformación del dataset, agregaremos (con una suma) la variable pax_total (cantidad total de pasajeros que usaron el molinete) por fecha, tipo de hora (pico y valle) y estación. Finalmente, como el dataset tiene pocas variables, vamos a agregar información proveniente de fuentes externas para enriquecer la información de la misma, desde la latitud y longitud de la estación hasta la comuna en la que se ubica, la densidad poblacional en su radio o el precio del boleto para ese mes. También se agregará información manualmente como, por ejemplo, si la estación es una cabecera de la línea, si conecta con una línea de tren o si es una combinación con otra línea de subte.

5.2 Modelos

Para predecir la cantidad de pasajeros utilizaremos el modelo del árbol de regresión, desarrollado por Leo Breiman (Breiman, 1984). Con el mismo buscaremos los mejores hiperparámetros mediante optimización bayesiana. De manera de poder evaluar la mejora en la predicción, compararemos nuestra performance con la de una regresión lineal múltiple. Adicionalmente, para generar un modelo base con el cuál evaluar en cuánto mejora la performance predictiva, se generará una regresión lineal múltiple con las funciones incorporadas en R.

5.3 Comentarios adicionales

Para evaluar la performance predictivo del árbol de decisión y compararla con los resultados arrojados por la regresión lineal múltiple, utilizaremos el *Root Mean Squared Error* (RMSE).

6. Fuentes

Alkhereibi, A. H., Wakjira, T. G., Kucuckvar, M., & Onat, N. C. (2023). Predictive Machine Learning Algorithms for Metro Ridership Based on Urban Land Use Policies in Support of Transit-Oriented Development. *Sustainability*, *15*(1718). https://doi.org/10.3390/su15021718

Beramendi, M., Gianotti, R., & María de los Ángeles, V. (2019). Un análisis exploratorio sobber las experiences de viajar en el subterráneo de Buenos Aires. *Psicodebate: psicologiá, cultura y sociedad*, *19*(2), 69-84. http://dx.doi.org/10.18682/pd.v19i2.960

Breiman, L. (1993). *Classification and Regression Trees*. Taylor & Francis.

Gutiérrez Puebla, J., Cardozo, O. D., & García Palomares, J. C. (2008). Modelos de demanda potencial de viajeros en redes de transporte público: aplicaciones en el Metro de Madrid. *Proyección*, *4*(7), 1-25. https://ri.conicet.gov.ar/handle/11336/60200

Starmer, J. (2019, August 20). *Regression Trees, Clearly Explained!!!* YouTube. Retrieved November 11, 2023, from https://www.youtube.com/watch?v=g9c66TUylZ4