Análisis de viajes en colectivo

Matías Lepore 152.457-4 | Matías Sica 152.245-0 | Federico Macias 152.635-2

Universidad Tecnológica Nacional (UTN), Facultad Regional Buenos Aires, Ingeniería Industrial, Buenos Aires, Argentina.

Abstracto

En este paper nos centramos en el análisis de los viajes en colectivos para descubrir tendencias y comportamientos a través de los últimos 6 años y generar un modelo capaz de predecir la cantidad de viajes que se realizaran por empresa en los próximos años.

Palabras claves

Colectivos, aprendizaje supervisado, regresión lineal, SVR, histograma.

1 INTRODUCCION

En el presente informe se desarrollará un análisis de los viajes realizados con el sistema único de boleto electrónico (SUBE) en colectivos en el área Metropolitana de Buenos Aires

El objetivo es aprender de los datos y poder predecir la cantidad de viajes que se realizaran en cada empresa de colectivos en el futuro.

2 DESCRIPCION DEL DATASET

El dataset utilizado es un dataset público obtenido de la página de Secretaria de Modernización de la Presidencia de la Nación llamado: "SUBE Operaciones de viaje por mes de Región Metropolitana de Buenos Aires" (Link: https://datos.gob.ar/dataset/spt-sube-operaciones-viaje-por-mes-region-metropolitana-buenos-aires).

El dataset contiene la cantidad de operaciones de viajes con SUBE en la Región Metropolitana de Buenos Aires discriminados por:

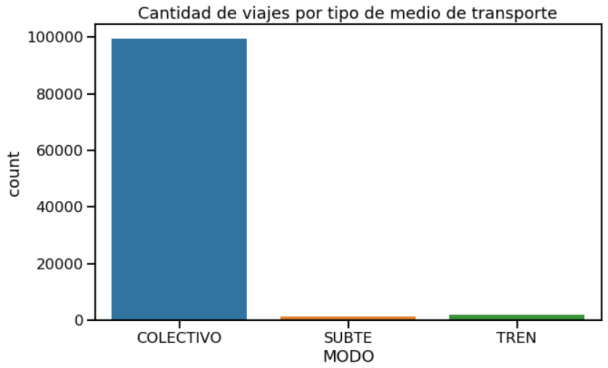
- Mes: desde 01/2013 hasta 06/2019
- Modo: colectivo, tren o subte
- Jurisdicción: nacional, provincial o municipal
- Grupo Tarifario
- Empresa
- Línea
- Tipo de Pasaje

En total contiene más de 102 mil samples.

3 ANALISIS EXPLORATORIO DE DATOS

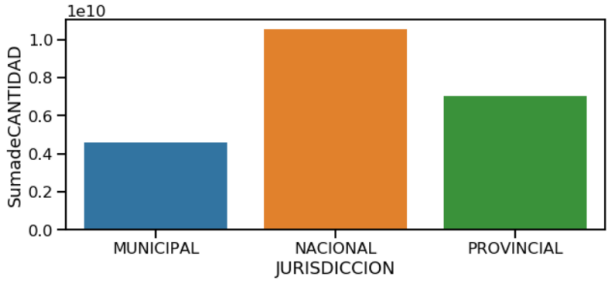
Para el análisis exploratorio de datos comenzamos analizando las diferentes features para entender el dataset y poder determinar cuáles de ellas aportaban mayor información.

En primer lugar, analizamos la variable "Modo" y realizamos un gráfico de barras, como se ve a continuación:



Notamos que el 97% de los datos se refiere a viajes en colectivos, por lo cual decidimos continuar analizando solo estos de aquí en adelante.

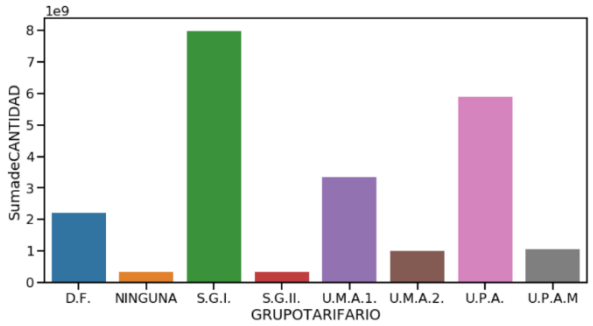
En segundo lugar, analizamos la variable "Jurisdicción":



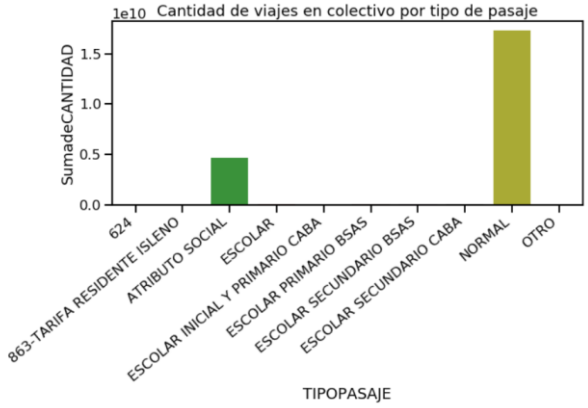
Como se puede observar la mayoria de los viajes en colectivo se realizan en lineas de jurisdiccion nacional. Dichas lineas son las comprendidas entre 0 y 200 que pertenecen a recorridos dentro de C.A.B.A.

Sin embargo, podemos ver que hay gran participacion de lineas que pertenecen a jurisdicciones provinciales y municipales. Por lo tanto, decidimos no descartar ninguna para realizar el analisis.

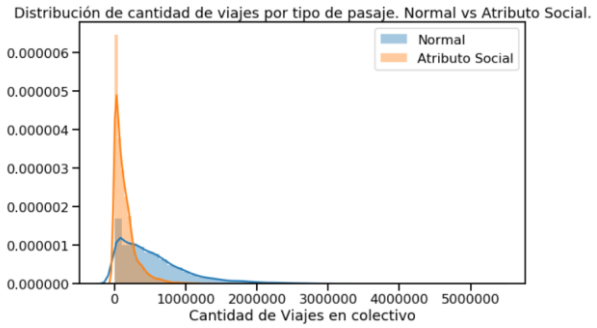
En tercer lugar analizamos la variable “Grupo Tarifario”. El grupo tarifario agrupa a las empresas de colectivos con sus lineas y kilometros recorridos según el Anexo III de la Resolucion RS 975 año 2018. (para mas informacion: https://www.transporte.gob.ar/UserFiles/boletin/ANEXOS-RESOLUCION-RS-975-2018-MTR/IF-RS-975-2018-MTRXV.pdf).



Luego analizamos la variable “Tipo de pasaje”:



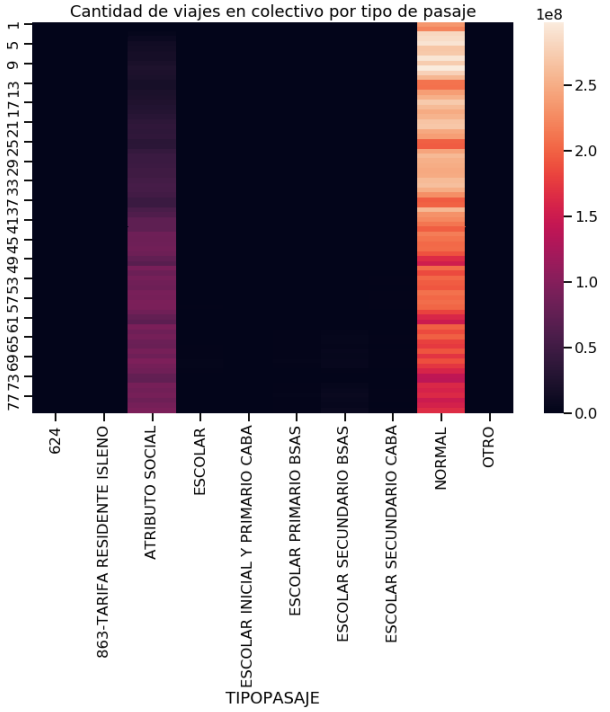
Como se puede observar en el gráfico, principalmente se utilizan dos tipos de pasajes: el normal y el atributo social. Es por esto que decidimos hacer un histograma para entender mejor cómo se comporta la variable.



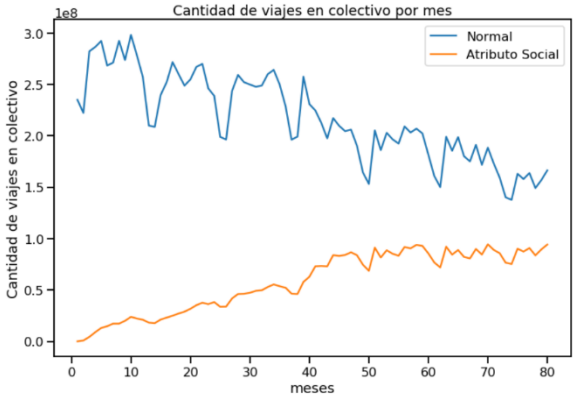
En relación al histograma, se ve claramente que la relación entre ambos tipos de pasajes se mantiene en la mayoría de las líneas de colectivo. Es decir, las líneas tendrán mayor cantidad de viajes con tipo de pasaje normal que con tipo de pasaje atributo social.

También nos pareció interesante analizar la cantidad de viajes de ambos tipos de pasajes a lo largo de los 80 meses que nos aporta el dataset.

A continuación, decidimos analizar cómo se comportaban los tipos de pasaje en el tiempo. Para esto realizamos un heatmap con la cantidad de viajes por tipo de pasaje



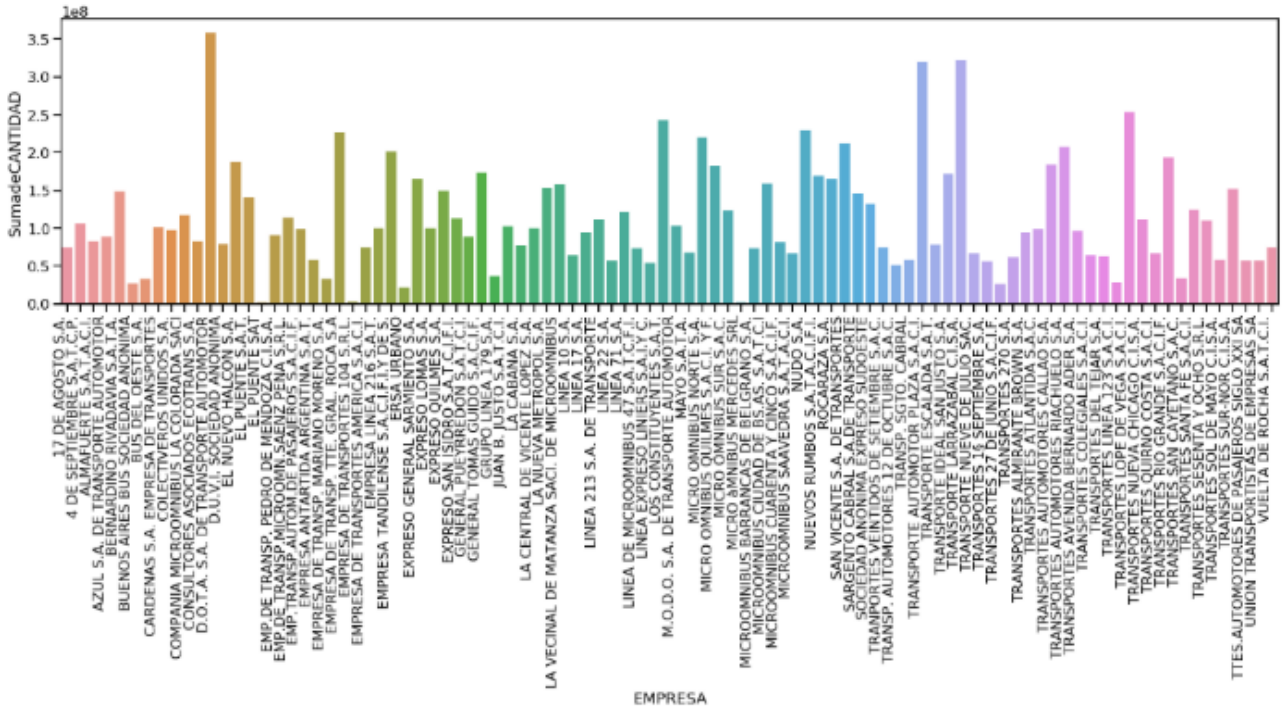
Como se puede observar hay una tendencia de aumento de boletos con atributo social y un decrecimiento de la normal, por lo que procedimos a realizar una serie de tiempo de ambos tipos de pasajes



Como conclusión de esto podemos decir que en los últimos años fue aumentando el tipo de pasaje atributo social, en detrimento del tipo de pasaje normal, manteniendo similar la cantidad total de viajes si se suman ambos tipos de pasajes.

Además, se observan picos descendientes en repetidas ocasiones en los meses de enero y febrero de los distintos años. Esto se puede deber principalmente al periodo vacacional escolar, lo que hace disminuir la cantidad de viajes.

Por otro lado, no se notan variaciones en la cantidad de viajes de atributo social dentro de la tendencia creciente notable que se observa.



Realizamos un conteo de viajes totales sin discriminar el tipo de pasaje para cada empresa, de esta manera pudimos evidenciar cuales son las empresas de colectivos que más se utilizan y sobre las que será más interesante aplicar el modelo de predicción de viajes en el futuro.

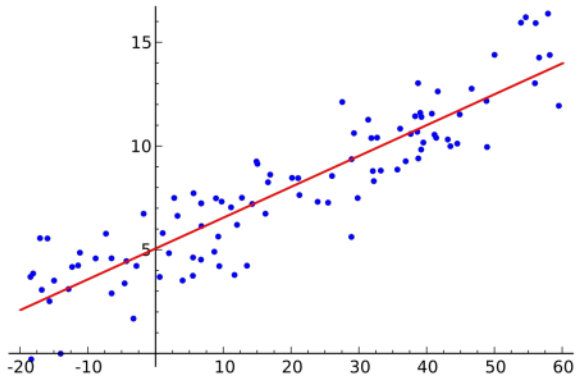
4 MATERIALES Y METODOS

Para realizar las predicciones por empresa utilizaremos regresiones:

- Regresion Lineal
- Support vector regression (SVR)

Regresión Lineal

El método regresión lineal es un modelo de machine learning supervisado, el cual consiste en determinar una recta que mejor se ajuste a los datos de entrada definidos.



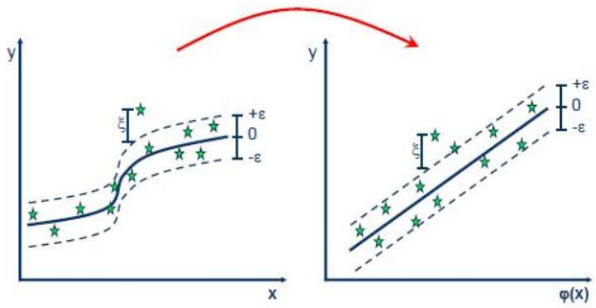
$$\widehat{y} = f\left(x, \beta\right)$$

De esta manera, el resultado obtenido es una etiqueta que será definida basándose en las muestras (x) y un coeficiente (β) que lo determina el modelo de la siguiente manera:

$$\min_{\beta} \|X_w - y\|^2 \longrightarrow \hat{\beta} = (X^T X)^{-1} X^T y$$

Support Vector Regression (SVR)

Este modelo al igual que la regresión lineal, es un modelo de machine learning supervisado, el cual busca construir una función lineal (hiperplano) que mejor se ajuste a los datos, definiendo ciertos márgenes donde se espera que los datos queden contenidos (ε).



El modelo busca maximizar dicho margen para que mayor cantidad de muestras caigan en su interior, pero sin maximizar el costo el cual define a épsilon.

$$C \sum_{n=1}^{N} \xi_n + 1/2 \|w\|^2$$

Evaluación de los modelos

Para poder medir los resultados obtenidos de los modelos recurriremos a:

- **MSE:** Error Cuadratico Medio

$$MSE = \frac{\sum (\hat{y}_t - y_t)^2}{n}$$

- **RMSE:** Raíz cuadrada del error cuadrático medio

$$RMSE = \sqrt{\frac{\sum (\hat{y}_t - y_t)^2}{n}}$$

- **MAE:** Media del error

$$MAE = \frac{|\sum (\hat{y}_t - y_t)|}{n}$$

5 RESULTADOS

Modelo regresión lineal para empresa Nuevos Rumbos

Al dataset original lo agrupamos por el nombre de la empresa Nuevos Rumbos dejando una pivot con el periodo dividido por mes y la suma de cantidad de viajes en cada uno.

Con las 80 samples que nos quedaron (Ene13' – Ago19') y nuestra feature de la cantidad de viajes dividimos nuestros datos en xtrain – ytrain y xtest – ytest con una relación de 80% train y 20% test.

Gráfico de la regresión lineal

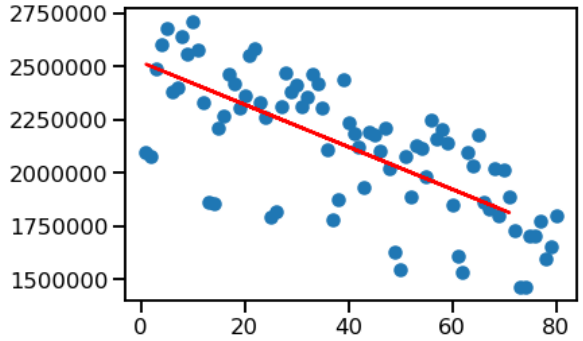
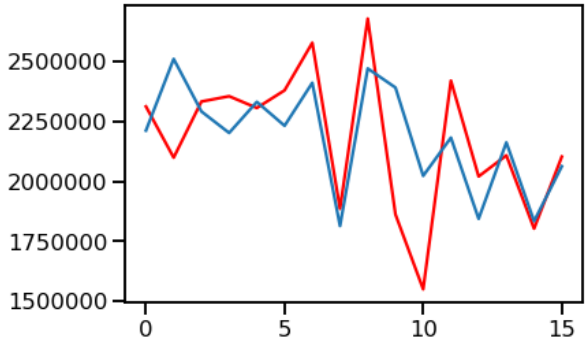


Grafico de Y predecida (azul) vs Y de test (rojo)



RMSE: 237.422,91

MAE: 179.427,10

MSE: 56.369.638.648,56

Modelo de regresión lineal para empresa DOTA

Este modelo se repite para la empresa que más viajes realiza (DOTA), para poder ver si los resultados del modelo siguen siendo buenos o no.

Gráfico de regresión lineal

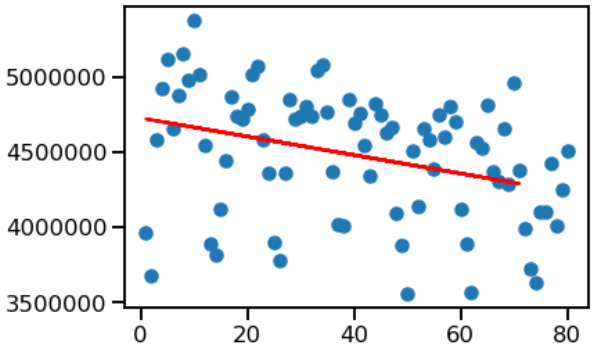
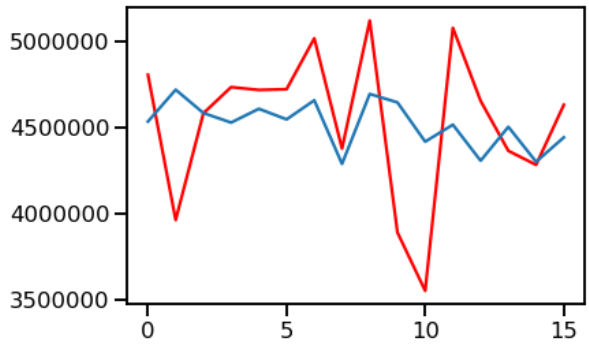


Grafico de Y predecida (azul) vs Y de test (rojo)

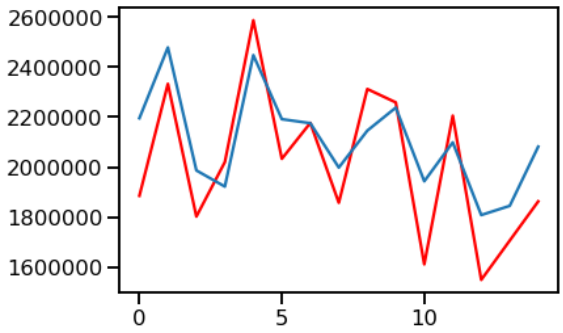


RMSE: 422.693,53

MAE: 329.260.14

MSE: 178.669.828.045,04

Modelo SVR para empresa Nuevos Rumbos

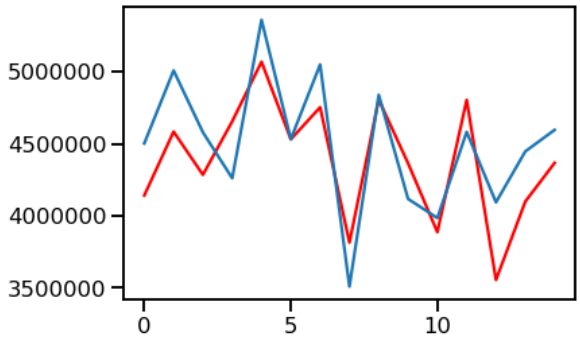


RMSE: 184.303,83

MAE: 161.385,05

MSE: 33.967.900.382,99

Modelo SVR para empresa DOTA



RMSE: 306.264,31

MAE: 272.570,18

MSE: 93.797.831.042,08

6 DISCUSIÓN Y CONCLUCION

A medida que fuimos analizando los datos pudimos descubrir características predominantes de los mismos. En primer lugar, el medio de transporte que predomina es el colectivo, luego pudimos encontrar diferentes tendencias, ya sea según el tipo de pasaje (atributo social, normal), según la jurisdicción de las diferentes líneas o según el grupo tarifario.

Finalmente fuimos adaptándonos a estas variables predominantes de las cuales teníamos más y mejores datos para poder trabajar con ellas. Alineándonos con el enfoque de machine learning, analizamos las empresas respecto de la cantidad de viajes registrados en los últimos 6 años.

Teniendo como base todo esto, nos pareció interesante, analizando la cantidad de viajes, poder predecir por empresa viajes futuros utilizando modelos de aprendizaje supervisado.

Como se pudo observar en los resultados es posible mediante los modelos predecir la cantidad de viajes para periodos futuros con un error aceptable (menor del 10%). De los dos métodos utilizados el que obtuvo mejores resultados es el SVR, el cual se ajusta mejor a los datos entrenados.

Con esta regresión es posible predecir para cualquier empresa de colectivos del dataset la cantidad de viajes que tendrán en el futuro analizando la tendencia que tienen los mismos.

Los resultados que se obtienen tienen gran utilidad para la empresa como forma de predecir tanto sus costos como ganancias y poder planificar mejor sus presupuestos.

De contar con más datos, los métodos se podrían replicar para subtes y trenes, llegando a predecir combinaciones entre los distintos medios de transporte para optimizar su interconexión y hasta mejorar infraestructura según las necesidades.

7 REFERENCIAS

- https://scikit-learn.org/stable/
- https://Stackoverflow.com
- https://github.com/clusterai