Predicción de subsidios al transporte público de pasajeros usando modelos de regresión

Análisis circunscrito a colectivos que operan en la RMBA

Melisa Breda y Miguel Mendez

1. Introducción

La red global del sistema de colectivos de la Región tiene una compleja organización y trama. En total, existen más de 300 líneas de diferentes jurisdicciones, aproximadamente 1.800 ramales, y una flota total de 9.931 vehículos.

De acuerdo a la Investigación de Transporte Urbano Público de Buenos Aires (Intrupuba[1]), se trata de un sistema de transporte con un alto grado de autonomía, dada su amplia cobertura y densidad de las redes, sumado a la frecuencia de los puntos de parada.

En relación a los usuarios de colectivos, la cantidad de pasajeros transportados por año en el periodo 2016-2018 fue de alrededor de 1.600 millones[2]. Según Intrupuba, las principales actividades que motivan los viajes en colectivo son el trabajo y el estudio. Y desde una óptica socioeconómica, el índice de nivel socioeconómico (INSE) muestra que el 80% aprox de hogares de los usuarios de colectivo se concentran en los niveles denominados Bajo Superior, Medio Inferior y Medio Típico; es decir que los principales usuarios de los colectivos pertenecen a clase media y media-baja.

Una particularidad de los colectivos es que, a diferencia del transporte ferroviario, se encuentran sometidos a diferentes regulaciones y fiscalizaciones jurisdiccionales (nacional, provincial y local ó municipal), de acuerdo a la inserción territorial de sus recorridos. Como consecuencia, existen diferentes normativas y niveles de exigencia y fiscalización de las empresas prestatarias y, a iguales condiciones de mercado, diferentes niveles de calidad de servicios ofrecidos, de tarifas cobradas y de rentabilidad.

Dado que se trata de un sistema de transporte vital para la población y que la crisis del 2001 perjudicó la ecuación económico-financiera de la empresas de transporte, el Estado intervino creando en el año 2002 una política de subsidios que funcionaba a través de dos mecanismos:

- subsidio directo en dinero a los operadores, conocido como Sistema Integrado de Transporte Automotor (SISTAU)
- subsidio en especies (litros de combustible – gasoil) a un precio diferencial en el mismo, de aproximadamente entre un tercio y un cuarto del precio de mercado, conocido como Cupo de Gasoil

Desde entonces, el Estado Nacional ha introducido diferentes modificaciones en la metodología de cálculo de subsidios que persiguieron diferentes objetivos, como por ejemplo:

- En el año 2006 se crea el Régimen de Compensaciones Complementarias (en adelante, RCC) al SISTAU, destinado a **compensar los incrementos de costos** incurridos por las empresas de

[1] La INTRUPUBA ha sido realizada por la Secretaría de Transporte de la Nación durante los años 2006 y 2007. Presenta los resultados correspondientes a los viajes realizados en todos los modos de Transporte Público (Ferrocarril, Subterráneo, Pre metro y Colectivo) https://www.argentina.gob.ar/transporte/dgppse/encuestas/estudios

[2] Estadísticas del Transporte Automotor https://www.argentina.gob.ar/transporte/cnrt/estadisticas-automotor

servicios de transporte público de pasajeros por automotor de carácter urbano y suburbano, que prestan servicios en el ámbito geográfico del AMBA, bajo Jurisdicción Nacional

- A partir de agosto de 2011 se **modifica el régimen de gasoil** para las empresas de transporte urbano: en lugar de asignar cupos de gasoil a través de declaraciones juradas, comenzaron a percibir un monto en dinero equivalente a la diferencia entre los litros de gasoil que hubieran demandado al precio fijado por la Secretaría de Transporte y el precio de mercado mayorista.

Las modificaciones más recientes han apuntado a **reconocer las diferencias de prestación de servicios en el AMBA** (ej: la Resolución 37 del año 2013 amplía la cantidad de Grupos Tarifarios) e incorporar nuevas compensaciones tarifarias asociadas a parámetros de demanda, tales como, atributo social, por recorrido superior a los 12 km y por usos global de tarifa (SUBE). Por último, en los últimos dos años se ha hecho un **uso más intensivo de la tecnología para incrementar el poder de control del Estado Nacional y perfeccionar el cálculo de los subsidios**. La resolución 937-E del 2017, por ejemplo, estableció que el monto de las compensaciones tarifarias se calcula tomando como base los kilómetros efectivamente verificados por línea, a través de la información que suministren los módulos de posicionamiento global (GPS) del Sistema Único de Boleto Electrónico (SUBE)[3]

El objetivo de este trabajo es identificar las principales variables que explican la asignación de subsidios a colectivos y evaluar si estamos frente a una asignación que contempla no solamente parámetros de oferta sino también de demanda. Nos preguntamos, concretamente, si la asignación de compensaciones tarifarias prioriza, en alguna medida, a aquellas empresas que transportan usuarios económicamente más desfavorecidos o no.

2. Descripción del Dataset

Construimos el dataset a partir de varias bases de datos obtenidas de la página de Datos Abiertos del Ministerio de Transporte de la Nación:

- Subsidios a colectivos de corta distancia Link
- Líneas de transporte de la RMBA Link
- Cantidad de operaciones de viajes por mes, línea y tipo de pasaje en RMBA Link
- Asignación S.I.S.T.A.U.[4] Link

Limitamos nuestro análisis a las líneas que operan en la RMBA[5] y para los años 2017, 2018 y 2019 (hasta el mes de Mayo)

A partir de este dataset, definimos que nuestra *variable a predecir* será el monto anual recibido en concepto de subsidio por empresa de colectivo. Este monto es la suma de dos componentes "tarifario" y "gasoil". Y definimos que las *variables independientes* a utilizar para predecir el monto anual de subsidio serán:

1. Viajes totales realizados por cada empresa de colectivo
2. Cantidad de líneas que posee cada empresa de colectivo
3. Cantidad de provincias en las que opera cada empresa de colectivo
4. Cantidad de municipios por los que pasa cada empresa de colectivo
5. Porcentaje de pasajeros transportados por empresa que abonan tarifa social[6]

[3] https://www.cronista.com/economiapolitica/Modifican-metodologia-para-los-subsidios-al-transporte-20171004-0030.html

[4] Sistema Integrado de Transporte Automotor

[5] La Región Metropolitana de Buenos Aires es la totalidad de los asentamientos urbanos, y sus respectivas áreas de influencia, integrados funcionalmente con el área urbana principal. Comprende una regionalización operativa y funcional que abarca a la Ciudad de Buenos Aires + 40 partidos de la Provincia de Buenos Aires. http://www.observatorioamba.org/planes-y-proyectos/rmba

[6] https://www.argentina.gob.ar/solicitar-la-tarifa-social-en-la-tarjeta-sube

Tabla I

Variables	Descripción
Variable a predecir	
Monto anual recibido en concepto de subsidio por empresa de colectivo	Subsidio recibido por CUIT, discriminado por dos componentes: tarifario y gasoil. Disponible para todas las provincias. Frecuencia del dato: mensual desde Ene/17 hasta May/19
Variables Independientes	
1. Viajes totales realizados por cada empresa de colectivo	Es la suma de los viajes realizados por empresa para los años 2017, 2018 y 2019 (sólo hasta Mayo) Frecuencia del dato: mensual
2. Cantidad de líneas que posee cada empresa de colectivo	Es el resultado de contar la cantidad de líneas asociadas a una empresa de colectivo.
3. Cantidad de provincias en las que opera cada empresa de colectivo	Es el resultado de contar la cantidad de provincias en las que opera una empresa de colectivo.
4. Cantidad de municipios por los que pasa cada empresa de colectivo	Es el resultado de contar la cantidad de municipios en los que opera una empresa de colectivo.
5. Porcentaje de pasajeros transportados por empresa que abonan tarifa social	Es el cociente A/B, donde: A = La suma de la cantidad de viajes realizados por una línea de colectivo para el tipo de pasaje "Atributo Social" para los años 2017, 2018 y 2019 (hasta Mayo) B = La suma de la cantidad de viajes realizados por una línea de colectivo para los años 2017, 2018 y 2019 (hasta Mayo)

El dataset Asignación S.I.S.T.A.U. lo utilizamos para establecer una conexión entre las empresas de colectivo y las líneas, dado que algunas variables -como el subsidio- están asignadas a empresas (identificadas por CUIT) mientras otras están asignadas a líneas de colectivos, como los viajes.

Un desafío que encontramos es que los nombres de las empresas estaban escritos de diferente manera en los diferentes datasets por lo que tuvimos que correr un proceso para asimilarlos. Para eso, usamos la librería de string matching **fuzzywuzzy** que utiliza la distancia de Levenshtein para calcular las diferencias entre secuencias y asigna un ratio de similitud entre dos strings. Sin embargo, esta buena herramienta no alcanzó para hacer toda la corrección de nombres, que se tuvieron que verificar manualmente.

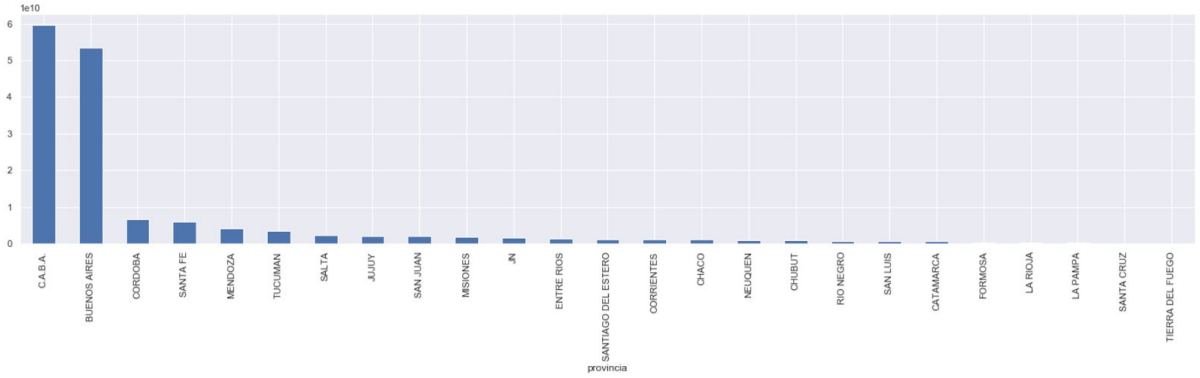
A partir de las bases de datos con las que trabajamos, arribamos a dos versiones de la variable Viajes totales realizados por cada empresa de colectivo.Optamos por utilizar la variable que se puede extraer de este dataset Cantidad de operaciones de viajes por mes, línea y tipo de pasaje en RMBA porque al presentar frecuencia mensualizada, nos permite filtrar los viajes hasta el mes de Mayo del 2019. Creemos que esta decisión le aporta más precisión al análisis, dado que los subsidios están disponibles hasta ese mes.

3. Análisis Exploratorio de Datos

Distribución de subsidios entre todas las provincias: vemos que CABA y la provincia de Buenos Aires (PBA de ahora en más) son las provincias que más subsidios reciben; circunscribir nuestro análisis a la RMBA no nos hace perder información significativa. CABA recibió en concepto de subsidios para el período en

análisis AR$59MM[7], la PBA AR$53MM, mientras que el resto de las provincias recibieron entre AR$6MM (Córdoba) y AR$8M (Tierra del Fuego). Esto implica que CABA y PBA explican el 74% de los subsidios otorgados en el período.

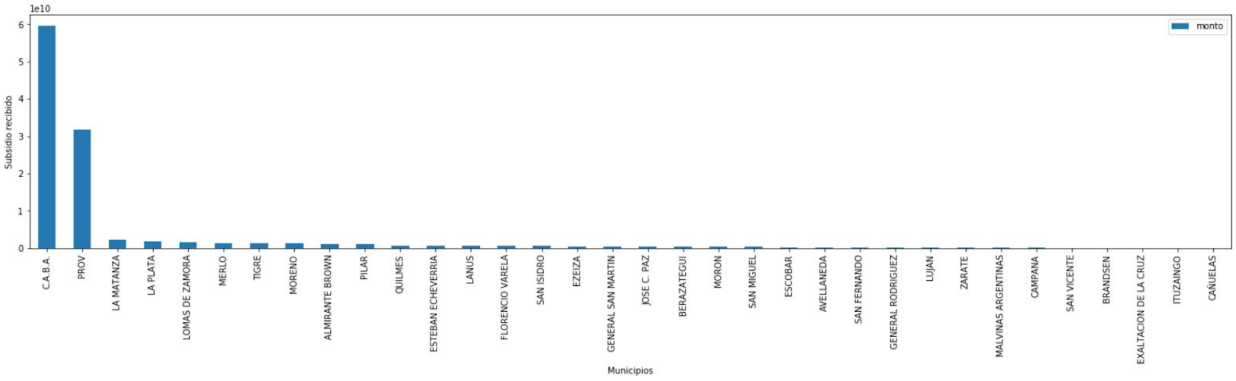
Gráfico I



Distribución de subsidios entre municipios de la RBMA: cuando miramos en detalle los municipios de la RMBA, vemos que la Ciudad es -por lejos- el municipio que más subsidios recibió. Esto es, por ejemplo, 27 veces lo percibido por La Matanza.

Lamentablemente muchos de los montos subsidiados no especifican el municipio en el cual se otorgaron, sino que figuran como "PROV", y esto nos evita conocer con un mejor detalle la distribución de los mismos. A pesar de esta imprecisión, decidimos conservar estas filas porque creemos que eliminarlas sería perder información valiosa.

Gráfico II



Evolución en el tiempo de los subsidios en la RMBA: cuando observamos la serie histórica distinguiendo entre los dos componentes (gasoil y tarifaria) vemos que el componente gasoil es significativamente menor al de tarifaria (oscila entre un 10% y un 35% del componente tarifaria) y que hay un máximo en la serie en Diciembre de 2018. Esto podría estar relacionado a una decisión política de recortar subsidios por parte del Estado Nacional, en el marco del cumplimiento del Consenso Fiscal, por el cual se trasladó parte de la responsabilidad de asignar subsidios a la Ciudad y la Provincia de Buenos Aires[8].

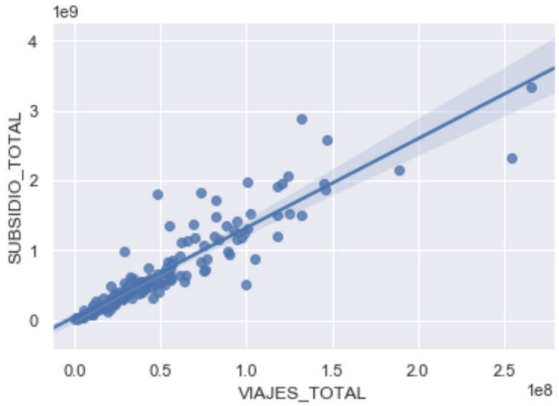
[7] MM indica miles de millones; M indica Millones

[8] https://www.cronista.com/economiapolitica/Nacion-dara-mas-subsidios-a-los-colectivos-en-el-GBA-para-que-no-aumenten-los-boletos-20190401-0053.html

Gráfico III

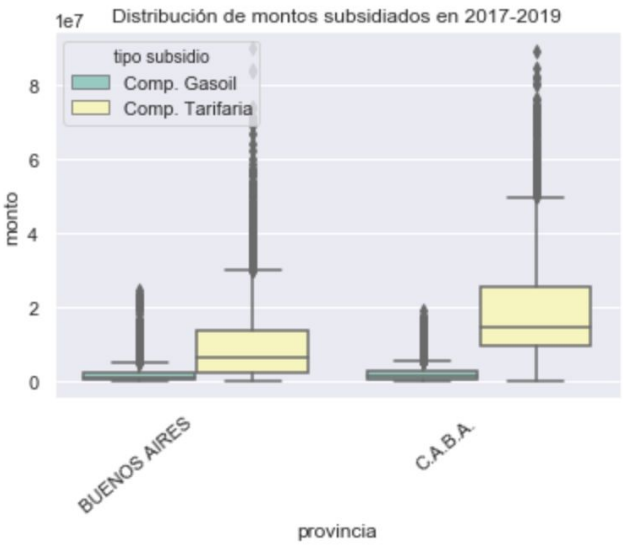
Relación entre viajes y total de subsidio: Si tomamos las variables Viajes SUBE Totales y Subsidios Totales, para el período analizado, podemos visualizar que existe una correlación lineal entre las mismas. Esto podría explicarse por el hecho de que el sistema SUBE permite un mejor control y direccionamiento de los subsidios por parte del estado.

Gráfico IV



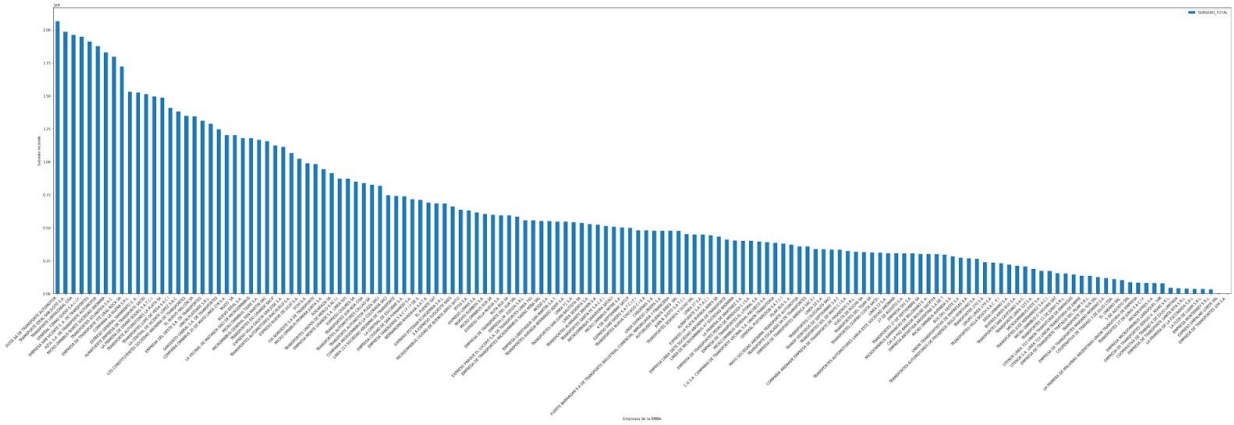
Distribución de subsidios por provincia y por tipo de componente: vemos que ambos componentes se comportan de manera bastante similar para ambas provincias. En relación al componente tarifaria, queda claro que la mediana de CABA es superior a la de la RMBA.

Gráfico V



Distribución de subsidios por empresas: sobre el dataset final, en el que trabajamos con 146 empresas, resulta que hay 30 empresas que concentraron aproximadamente el 50% de los subsidios otorgados en el período. Este grupo de empresas que más subsidios recibieron está encabezado por la empresa DOTA S.A[9]
.

Gráfico VI



4. Teoría y Modelos

Dividimos nuestro dataset en 90% train y 10% test. Autoescalamos los datos de manera tal que tengan una distribución con media=0 y desvío estándar=1. También eliminamos outliers usando como corte el cuantil 98,5

Aplicamos tres modelos de regresión:

i. *Linear Regression:* el objetivo es construir una función lineal que minimice el error de predicción. Para eso definimos la siguiente función de costo, que mide cuán inaccurate son las predicciones de nuestro modelo:

$$Cost = \frac{\sum_1^n((\beta_1 x_i + \beta_0) - y_i))^2}{2 * n}$$

ii. *Support Vector Machines Regression (SVR):* el objetivo del SVR es ajustar el error de predicción dentro de un cierto umbral (epsilon). Corrimos el modelo con dos tipos de kernel (lineal y gaussiano). Si bien SVR con kernel gaussiano provee la flexibilidad de generar una frontera no lineal, para nuestro modelo parece más acertado usar una frontera lineal.

iii. *KNN Regression*: en este modelo la variable objetivo es predicha por interpolación local de los valores de la variable objetivo asociados a los K vecinos cercanos en el training set. El hiper-parámetro a definir es K.

Los indicadores que usamos para evaluar la performance de cada modelo son:

- *MAE - Error Absoluto Medio*

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

- *MSE - Error Cuadrático Medio*

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

[9] https://www.lanacion.com.ar/economia/un-solo-grupo-empresario-controla-la-mitad-de-los-colectivos-del-amba-nid2292502

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

- *RMSE - Raíz del Error Cuadrático Medio*
- *R^2 :* mide la proporción de la varianza total de la variable explicada por la regresión

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

$$\text{MSE(baseline)} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

donde

El *MSE* es preferible al MAE pues "castiga" errores grandes. Adicionalmente, el *RMSE* es preferible al *MSE* porque es más fácilmente interpretable, dado que el resultado está expresado en la misma unidad que nuestra variable a predecir "y".

5. Seteo de Hiperparámetros

Todos los hiper-parámetros fueron calibrados usando el train set. Usamos accuracy como la medida de performance de los mismos. Estos fueron los parámetros elegidos:

Tabla II

Modelo	Hiperparámetros
Support Vector Machines Regression	Kernel=lineal C= 100
KNN Regression	K=20 Distancia=euclídea

6. Resultados / Performance

La Tabla III muestra los resultados para los diferentes modelos aplicados. Los mejores resultados son los señalados en azul.

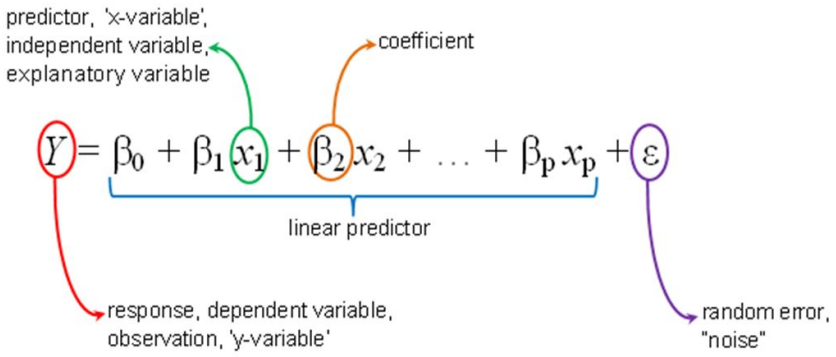
Tabla III

	MAE	MSE	RMSE	R2
Linear Regression	133.548.593	3.55e+16	188.487529	Training score: 0.71 Test score: 0.85
Support Vector Machines Regression (SVR)	332.540.112	2.33e+17	483.131.972	Training score: -0.07 Test score: -0.01
KNN Regression	131.060.732	4.28e+16	206.786.963	Training score: 1.00 Test score: 0.82

Los resultados nos dicen que:

- El mejor modelo de predicción es el de Regresión Lineal.
- El R^2 de nuestro modelo es aceptable por lo que podemos decir que la bondad de ajuste de nuestro modelo a la variable que estamos buscando explicar es alta.
- Todas las variables que utilizamos tienen poder de predicción, dado que corrimos un ejercicio de feature selection con Lasso (alpha=0.01) y todos los coeficientes asociados a nuestras variables resultaron diferentes de cero.

Recordando que en nuestro caso estamos utilizando Multiple Linear Regression, por tratarse de un modelo que utiliza más de una feature para predecir la variable dependiente y.



[10]

Exploramos cuáles son los pesos de nuestras variables en nuestro modelo. Los resultados se muestran en la Tabla IV:

Tabla IV

Modelo/Coeficientes	**Cantidad Líneas**	**Cantidad Municipio**	**Cantidad Provincias**	**Viajes Totales**	**% Atributo Social**
Linear Regression	4.47529200e+07	6.91676306e+07	7.48578701e+07	3.62082770e+08	-1.06240820e+08

Nuestro modelo, por lo tanto indica que:

- La *variable dependiente subsidios* está **directamente** relacionada con todas las variables que hemos considerado como proxys del costo de operación de las empresas (cantidad de líneas, cantidad de municipios y de provincias que recorre y la cantidad de pasajeros que traslada) Esto nos permitiría afirmar, de manera muy generalizada, que la metodología de cálculo y asignación de subsidios **efectivamente contempla parámetros de oferta**.
- La variable cantidad de pasajeros podría tomarse como una proxy de demanda y de oferta al mismo tiempo, dado que por ejemplo cuando una empresa traslada una mayor cantidad de pasajeros, esto repercute directamente en su operación, obligando a -quizás- comprar más unidades o aumentar la frecuencia del servicio.
- Por último, y quizás lo más interesante a los efectos de nuestro trabajo, la *variable dependiente subsidios* está **inversamente** relacionada con la variable que utilizamos como proxy de demanda (porcentaje de pasajes que corresponden a categoría Atributo Social) Esto nos permitiría afirmar, de manera muy generalizada, que la metodología de cálculo y asignación de subsidios **no tiene en consideración parámetros de demanda de los sectores más desfavorecidos (jubilados / pensionados / beneficiarios AUH / ex-combatientes de Malvinas, etc)**. Es más, por tratarse de un coeficiente negativo, podríamos afirmar que las empresas que trasladan un mayor porcentaje de este tipo de usuarios, reciben menos subsidios. Las razones detrás de este fenómeno exceden este trabajo.

[10] https://hackernoon.com/an-intuitive-perspective-to-linear-regression-7dc566b2c14c

7. Conclusiones y Trabajo Futuro

La principal conclusión y sugerencia de trabajo futuro está relacionada con la calidad de los datos. Consideramos que, si bien los resultados son intuitivos, podríamos generar un mejor modelo con mejor poder predictivo a partir de trabajar con features que capturen mejor las variables que están detrás de la formulación de subsidios. Los datos disponibles en la web de Gobierno Abierto del Ministerio de Transporte, que fueron nuestra fuente de información, definitivamente podrían ganar en claridad, orden y riqueza.

En segundo lugar, en el camino hemos hecho varios supuestos y tomado diversas decisiones acerca de las variables que utilizamos, como por ejemplo, eliminar outliers o filas donde los valores eran NANs. Es importante tener en cuenta que los resultados de este trabajo, están condicionados por esas decisiones.

En tercer lugar, los resultados en relación a la relación entre las variables independientes y nuestra variable a predecir parecen ir muy en línea con lo que diversas publicaciones y estudios sugieren acerca del esquema de subsidios a los colectivos: los subsidios siguen sin atender aspectos relacionados con el tipo de usuarios que los utilizan. Más allá de las modificaciones introducidas recientemente por el Estado Nacional a los fines de incorporar conceptos calculados en función de parámetros de demanda, los resultados de nuestro trabajo no permiten confirmar esto último. Este aspecto amerita, sin lugar a dudas, más exploración y podría ser objeto de futuros trabajos.

Por último, futuros trabajos podrían intentar mejorar la performance de nuestro modelo aplicando otros métodos como por ejemplo Lasso Regularized Linear Regression para reducir el riesgo de overfitting y detectar si algunas variables están siendo asignadas demasiado peso.

8. Referencias

- Datos Abiertos del Ministerio de Transporte: https://servicios.transporte.gob.ar/gobierno_abierto/
- Intrupuba: https://www.argentina.gob.ar/transporte/dgppse/encuestas/estudios
- Estadísticas del Transporte Automotor: https://www.argentina.gob.ar/transporte/cnrt/estadisticas-automotor
- https://www.lanacion.com.ar/economia/como-se-componen-los-subsidios-nid1482402
- https://repositorio.utdt.edu/bitstream/handle/utdt/2138/MEU_2015_Fernandez.pdf?sequence=1&isAllowed=y
- https://www.argentina.gob.ar/solicitar-la-tarifa-social-en-la-tarjeta-sube