PROYECTO APLICADO

Presentado a

PONTIFICIA UNIVERSIDAD JAVERIANA

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

Autores:
Juan Manuel Hincapié Martínez
Giovanni Alexander Montoya Garzón
Juan Mauricio Moreno Chanchay
Gustavo Adolfo Ochoa Blanco

Prueba de concepto para Generación Móvil

2021

Contenido

Lista de figuras

Lista de tablas

1. Entendimiento del negocio

Generación Móvil es una compañía dedicada al desarrollo e implementación de las ZER (Zonas de Establecimiento Regulado) en Colombia. Este trabajo lo vienen realizando por varios años en conjunto con las entidades gubernamentales de cada municipio donde se encuentran estas zonas. Cada municipio es el encargado de reglamentar las ZER y en el país se encuentra en ciudades principales como Medellín y Cali.

El objetivo principal del programa ZER es mejorar la movilidad urbana, recuperando de manera óptima el espacio público y la desincentivación a la utilización de la calle como zona de parqueo; además de motivar a las personas a utilizar medios de transporte más ecológicos y sostenibles, que el uso de vehículos. Las ZER son zonas demarcadas en las vías del municipio, que permiten el estacionamiento de vehículos con pago de tarifa por hora o fracción, la cual es recolectada por un promotor, este por lo general tiene varias zonas a cargo dentro de un municipio, las cuales recorre durante el transcurso del día *n* cantidad de veces.

Para efectos de este proyecto, se analizará las ZER del municipio de Caldas – Antioquia, el cual reglamentó las Zonas de Estacionamiento Regulado en el decreto 144 en septiembre del 2017, con 19 zonas de acuerdo con puntos de referencia y una capacidad de 202 celdas para carros y 254 celdas para motocicletas. Un promotor tiene asignadas entre 4 y 6 zonas a su cargo, cada zona tiene mínimo una cuadra de celdas demarcadas, el horario es de 7 am a 7 pm y se trabaja de lunes a sábado.

Actualmente la concesión encargada de las ZER en el municipio de Caldas presenta perdidas en su ejercicio contable, ya que los costos, entre ellos la nómina de promotores, comparados a los ingresos percibidos por el uso de las celdas no compensan ni generan excedentes para garantizar estabilidad financiera de la actividad. Generación Móvil ha estado trabajando en las posibles causas o efectos que estén impactando negativamente en la utilidad y ha encontrado que la emisión de un tiquete en ocasiones no se genera al momento de que el vehículo llega a la celda, si no cuando el promotor llega a la zona y percibe el ingreso de un nuevo vehículo; por esto hay un desfase de tiempo entre el inicio real y transaccional del evento. De igual manera, al momento de finalizar un evento, puede que no se encuentre el promotor y los propietarios de los vehículos se retiren de la ZER sin pagar.

Siendo así, el objetivo de negocio principal es apoyar en la toma de decisiones en cuanto a la programación de rutas de los promotores, de manera que se minimicen los problemas mencionados en el párrafo anterior: tiempo muerto entre llegada real de un vehículo a la ZER e inicio de la transacción y retiro de la ZER sin pagar. Para llevar a cabo el análisis, se cuenta con cuatro consultores especializados en analítica de datos, plataformas como anaconda, Google Collaboratory y Power BI, datos transaccionales, de clima y otros brindados por el cliente. Para lograr el objetivo de negocio planteado, como objetivos de analítica de datos se proponen: predecir de manera

oportuna los ingresos y salidas de las ZER, predecir la duración de un vehículo en la ZER y, por último, identificar los clientes que tienen probabilidad de retirarse de la ZER sin pagar.

El plan para lograr los objetivos de analítica de datos se basará en 4 pasos generales: entendimiento y limpieza de los datos suministrados por el cliente y su integración con fuentes de datos externas, desarrollo de series de tiempo a partir de los datos suministrados por el cliente para realizar predicciones basadas en tiempo, desarrollo de modelo de aprendizaje supervisado para predecir duración de estadía de un vehículo en la ZER y desarrollo de modelo de clasificación binaria para identificar los clientes que incurren en fraude en el uso de la ZER. Con estos cuatro pasos mencionados, el cliente estará en posición de optimizar las rutas de los promotores.

2. Exploración de datos

Generación Móvil proporcionó una base de datos transaccionales, dividida en dos tipos de registros: los tiquetes pagos y los no pagos. El grupo de tiquetes pagos tiene 501530 registros y el grupo de tiquetes no pagos consta de 32661 registros. La base de datos tiene las siguientes columnas:

- *ID del registro*: número único identificador de la transacción.
- *Código Zona*: Identificación de la ZER. La base de datos contiene 20 zonas diferentes.
- *Nombre del promotor*: Nombre de la persona que trabaja para la ZER emitiendo y cobrando tiquetes.
- *Placa del vehículo*: Identificador del vehículo
- *Tipo de vehículo*: Automóviles o motos.
- *Tipo de ingreso*: Pago al ingreso, normal o prepago.
- *Fecha/hora de ingreso*: Estampa de tiempo del inicio de la transacción.
- *Tipo de cobro*: Tiquete pago, cartera recuperada o cobro adicional. Cartera recuperada se refiere a un cliente que no pagó el tiquete cuando se retiró de la ZER pero tiempo después volvió y pagó el tiquete pendiente; cobro adicional se causa para algunos clientes con tarifa especial.
- *Tipo de pago*: Muy relacionada con el tipo de tarifa, los tipos son normal, tiquete no pago, salida de otra zona, tiempo de gracia y cobro prepago.
- *Tarifa*: Existe la tarifa normal denominada ZER CALDAS y adicionalmente hay tres tarifas especiales que son prepago, pospago y convenio general. Cada una de estas tarifas se diferencian para carros y motos.
- *Valor*: Valor en COP de la transacción.
- *Número de factura*: Identificador de la factura.
- *Fecha/hora de factura*: Estampa de tiempo de la emisión de la factura.
- *Operador salida*: Nombre del promotor en sitio al momento de la salida del vehículo.
- *Fecha/hora de salida*: Estampa de tiempo del cierre de la transacción.
- *Tiempo total:* Tiempo total de la estadía en horas, minutos y segundos.
- *Número de personal:* Número identificador del promotor.
- *Posición de la celda:* Número identificador de la celda de parqueo.

A continuación, se presenta un análisis exploratorio inicial de los datos.

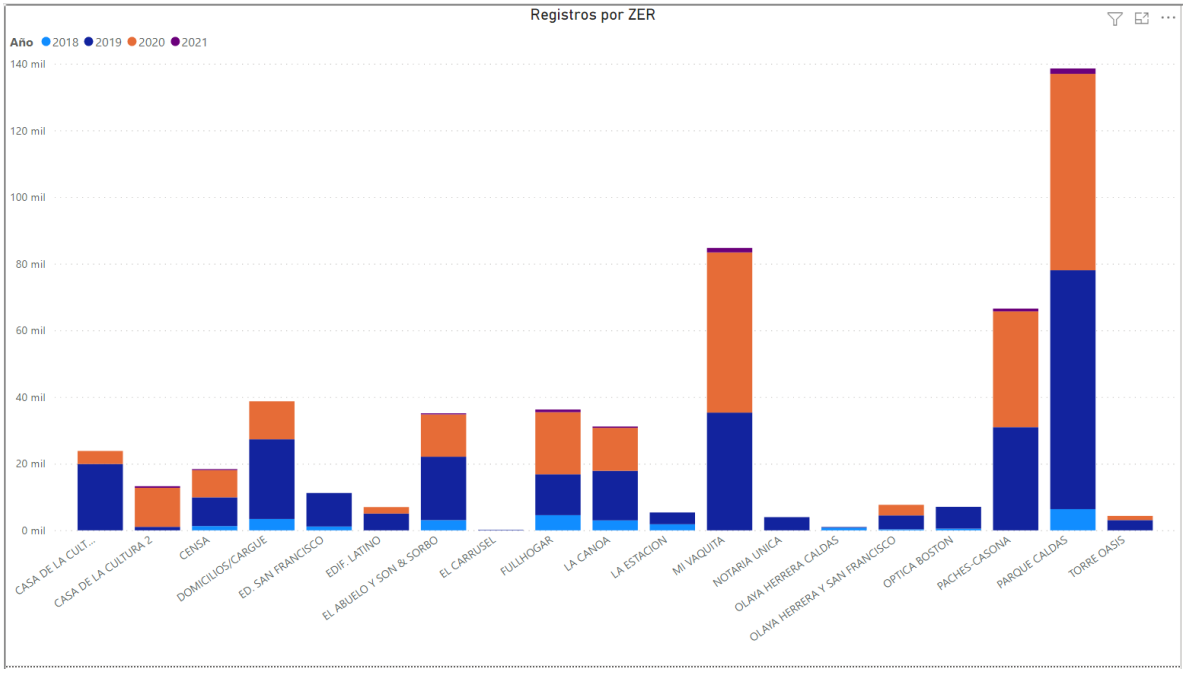

Figura 1. Distribución de registros de la base transaccional por ZER.

En la Figura 1 se presenta cómo se distribuyen los registros de la base de datos transaccional entre las diferentes ZER. En total hay 19 zonas, de las cuales, las más utilizadas son Parque Caldas, Mi Vaquita y Paches-Casona. Adicionalmente, se logra un primer acercamiento al impacto del efecto COVID19. Para zonas como Edificio San Francisco, La Estación, Notaría Única, Olaya Herrera Caldas y Óptica Boston se observa cesación de registros para el 2020, lo que indica un posible cierre temporal o permanente de la zona. Por otro lado, las zonas que sí presentan registros en el 2020, se observa un decremento respecto al año 2019. En la Figura 3 se presenta con mayor claridad el impacto del efecto COVID19.

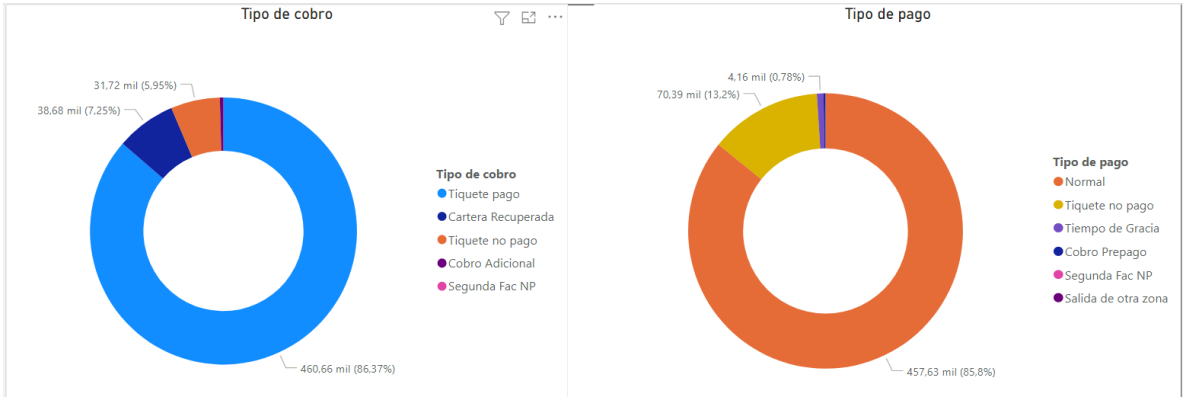

Figura 2. Relación entre variables Tipo de cobro y Tipo de pago.

En la Figura 2 se observan las diferentes clases de tipos de cobro y pago para las ZER. Se evidencia la correlación entre los tiquetes que se generan normalmente y son pagados por los clientes. Adicional se puede ver la porción de la cartera recuperada, la cual se genera por los tiquetes no pagados.

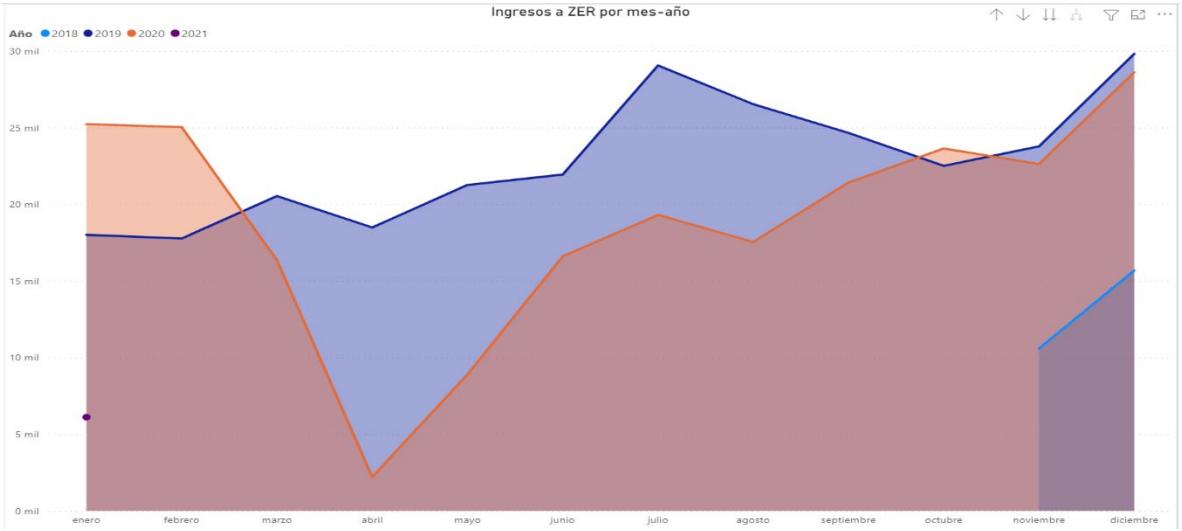

Figura 3. Evolución de ingresos a las ZER a través del tiempo. Granularidad mensual.

En la Figura 3, se evidencia el efecto COVID19 en la ocupación de las ZER, ya que a partir del mes de marzo de 2020 (declaración de emergencia sanitaria en Colombia e inicio de cuarentena general) se ve el bajo nivel de ocupación y su recuperación poco a poco desde el mes de junio. Sin embargo, es interesante que a pesar del efecto COVID19, el año 2020 terminó de manera similar al año anterior (2019), se podría decir que el municipio se ha estado movilizando de manera similar.

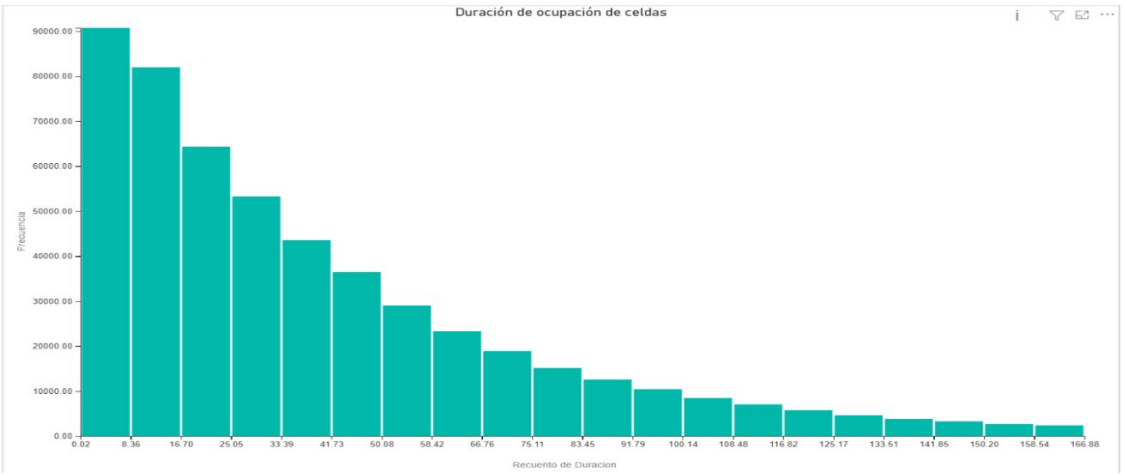

Figura 4. Distribución de la duración de ocupación.

En la Figura 4 se puede observar como la mayoría de los usuarios de las ZER ocupan esos espacios por un periodo corto de 0 a 8 minutos aproximadamente, esto se debe al tipo de negocio ya que no se especializa en usuarios de largo plazo. En cambio, su principal foco son clientes que buscan estacionarse por pocos minutos para realizar algún tipo de diligencia y retirarse lo más pronto posible. El comportamiento descrito, hace que la distribución de la duración tenga una asimetría positiva, marcada por un modelo de negocio *long-tail*.

Por último, en la Figura 5 se identifica la cantidad de placas únicas presentes en la base de datos: 98.967 placas diferentes, pero de éstas 50.858 sólo tienen una transacción. La que más transacciones tiene cuenta con 737 seguida de una placa con 329 transacciones (las que siguen no

están tan alejadas de la segunda). En la vista de la Figura 5 se hace un *zoom-in* a la Figura 4, y se observa que la duración más común es de 4 minutos, con un total de 5116 transacciones con esta duración. Sin embargo, el promedio de duración de toda la base de datos es de 45.26 minutos.

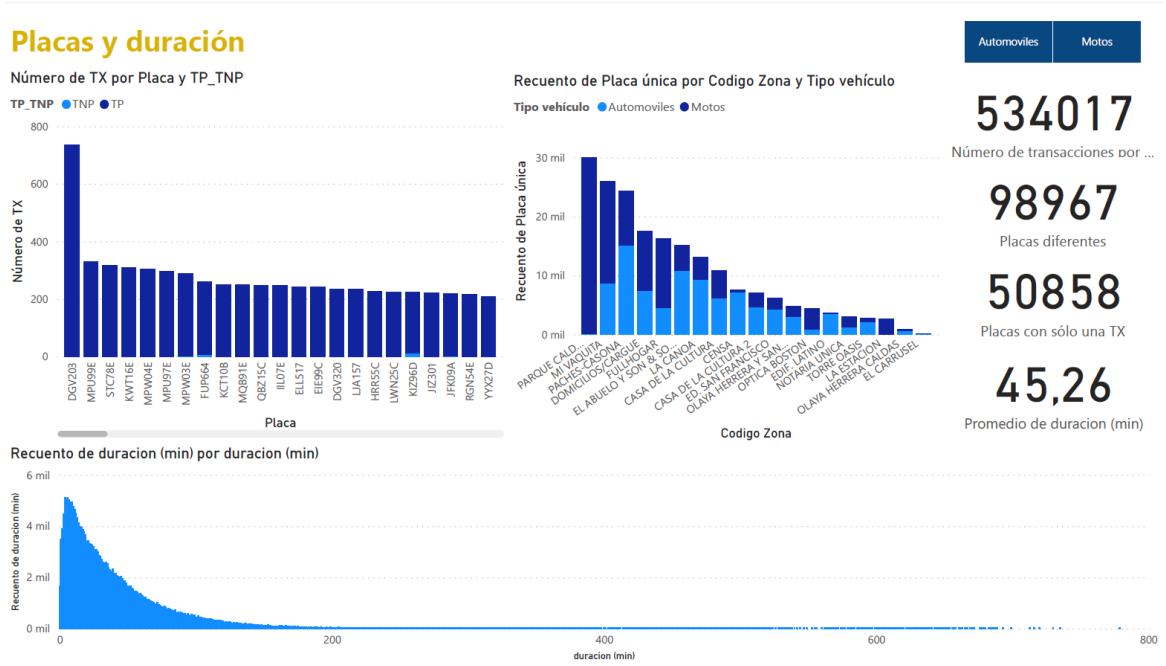

Figura 5. Análisis de duración y placas.

3. Preparación de datos

Una vez explorada la base de datos inicial, entregada por el cliente, se procede a hacer las respectivas transformaciones para desarrollar los modelos analíticos. Como se explicó al final de la sección 1 del presente documento, se desarrollarán 3 modelos de analítica diferentes para abordar los 3 objetivos de analítica: predecir cuántos ingresos y salidas habrá para los siguientes 5 minutos en cada zona, predecir la duración en la celda de parqueo de un nuevo cliente y predecir si un cliente pagará o no pagará el servicio de parqueo. La Figura 1 se muestra que la base de datos inicial contiene 19 zonas de parqueo, sin embargo, en conversaciones sostenidas con el cliente, en la actualidad existen 12 zonas. Inicialmente se unen los listados de transacciones de tiquetes pagos y no pagos. Después, la primera modificación realizada entonces es la reclasificación de las zonas como se muestra en la Tabla 1.

La segunda modificación fue remover aquellas transacciones con problemas claros de calidad de datos haciendo las siguientes validaciones: todas las transacciones inician y finalizan el mismo día y deben tener una hora de finalización mayor a la de inicio. Con estas revisiones de calidad se eliminan 191 transacciones de la base de datos, que corresponde al 0.03% de todas las transacciones.

Después de actualizar los nombres de las zonas de parqueo y hacer la verificación de calidad de datos según el conocimiento de negocio se presenta la preparación de datos realizada para obtener las bases a utilizar en cada uno de los modelos.

Tabla 1. Reclasificación de zonas de parqueo.

Zonas BD original	Zonas modificadas
EL ABUELO Y SON & SORBO	EL ABUELO Y SON & SORBO
CENSA	CENSA
LA CANOA	LA CANOA
FULLHOGAR	FULLHOGAR
PARQUE CALDAS	PARQUE CALDAS
DOMICILIOS/CARGUE	DOMICILIOS/CARGUE
MI VAQUITA	MI VAQUITA
CASA DE LA CULTURA	CASA DE LA CULTURA_mod
CASA DE LA CULTURA 2	
EDIF. LATINO	LATINO Y CARRUSEL
EL CARRUSEL	
NOTARIA UNICA	CASONA Y NOTARIA
PACHES-CASONA	
OPTICA BOSTON	BOSTON Y OASIS
TORRE OASIS	
ED. SAN FRANCISCO	LOCERIA
LA ESTACION	
OLAYA HERRERA CALDAS	
OLAYA HERRERA Y SAN FRANCISCO	

3.1. Preparación de datos para predicción de ingresos y salidas en los próximos 5 minutos para cada zona

Este objetivo se abordará mediante dos tipos de modelos analíticos: series de tiempo y caracterización mediante estadística descriptiva. Sabiendo esto, lo principal es transformar la base de datos transaccional a series de tiempo con frecuencia 5-minutal. Teniendo en cuenta que el objetivo es lograr predecir la cantidad de ingresos y salidas en cada una de las zonas y discriminando por tipo de vehículo será necesario desarrollar 48 series de datos en total: 12 zonas, 2 tipos de vehículos, ingreso o salida.

Para obtener las 48 series de tiempo, se creó un *dataframe* en el que las filas corresponden a las ventanas 5-minutales y las columnas a cada serie de tiempo identificadas de la siguiente manera: "ZONA:TIPO_VEHÍCULO:INGRESO/SALIDA". Para el cálculo de ingresos y salidas se utilizaron las variables originales *Fecha/hora ingreso* y *Fecha/hora salida* respectivamente. Sobre estas series de tiempo se hicieron validaciones de calidad, confirmando que la cantidad de ingresos de cada tipo de vehículo en cada zona fuera igual a la cantidad de salidas.

3.2. Preparación de datos para estimación de duración

Para el modelo de aprendizaje supervisado de estimación de duración, se realizó un análisis de incidencia de variables sobre la variable objetivo: en este caso la duración de la transacción. Las variables como *nombre de promotor* y *placa de vehículo* tienen una alta cardinalidad por lo que no influyen en la variable objetivo. Por otro lado, las variables *número de personal* y *posición de la celda* tienen el mismo dato en todas las transacciones por lo que también son eliminadas de la base de

datos. También, las variables de *tiquete pago o no* pago, *fecha/hora de salida* y *valor* se eliminan pues por la naturaleza del problema a resolver no estarían disponibles al momento de realizar una estimación de duración. Al final, se utilizan las variables *código zona*, *tipo de vehículo*, *tipo de ingreso*, *tipo de pago*, *tipo de cobro*, *tarifa* y *duración* (variable objetivo) para el desarrollo del modelo.

Adicionalmente, se categorizó las horas del día en mañana (de 7 a.m a 12 m), medio día (de 12 m a 2 p.m) y tarde (de 2 p.m a 7 p.m). Combinando las placas, las zonas y la jornada de trabajo se desarrolló un cubo de información, donde cada elemento corresponde al valor de duración promedio para una placa, una zona y una jornada específica. Adicional de los elementos particulares del cubo, también es posible agregar por placa, para obtener el promedio general de duración de la placa; por zona, para obtener el promedio general de duración de la zona; o por jornada. A continuación, en la Figura 6 se presenta gráficamente el cubo generado. Cada uno de los datos del cubo se utiliza para caracterizar cada una de las transacciones.

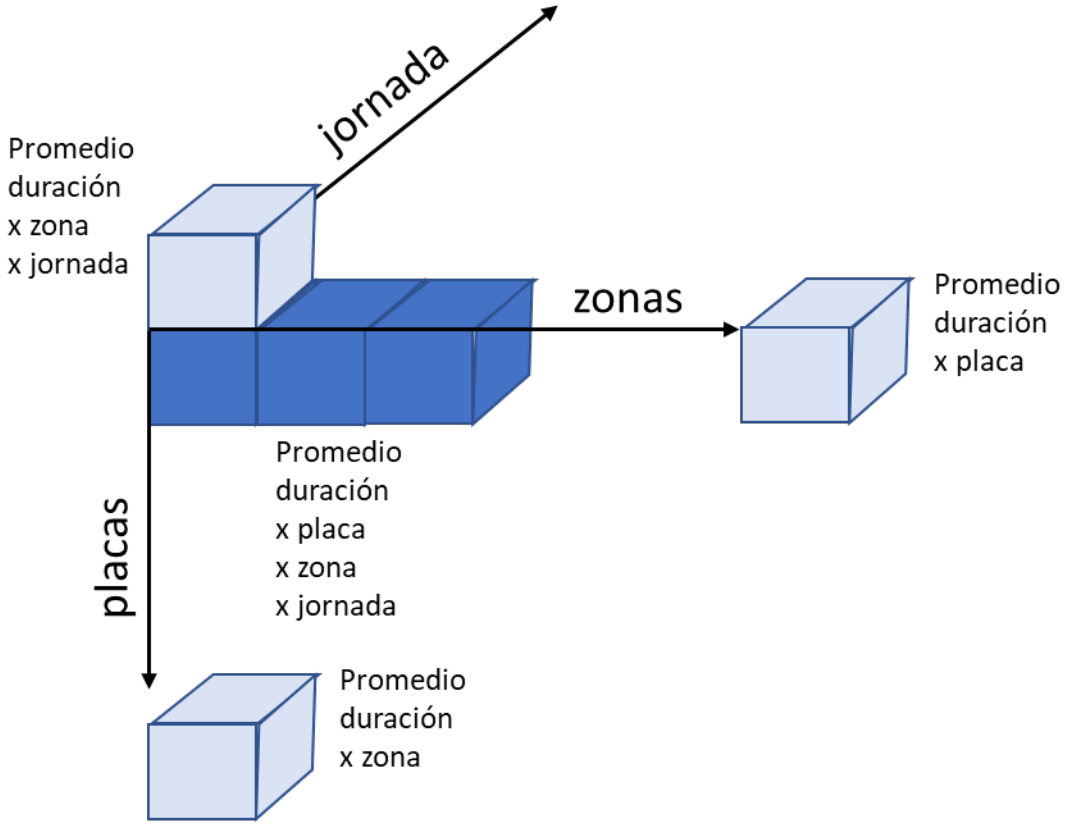

Figura 6. Cubo de caracterización de placas y zona respecto a duración de transacciones. Elaboración propia.

Con este cubo, las variables utilizadas para desarrollar los modelos de regresión de duración de transacción son el promedio de duración por zona y jornada, el promedio de duración por placa, la hora de ingreso (como número entero de hora, no confundir con estampa de tiempo), el día de la semana de la transacción y el score de fraude – que será explicado en la siguiente sección –.

3.3. Preparación de datos para modelo de fraude

Este modelo tiene como objetivo explicar el comportamiento del cliente ante una transacción, es decir, modelar si esa transacción será paga o no paga. Para el desarrollo de este modelo se resalta que para la categorización de los clientes es necesario empezar con la definición de un *score* capaz de describir o capturar el comportamiento transaccional de éstos. En este caso, cada cliente está identificado por placa dentro de la base de datos transaccional.

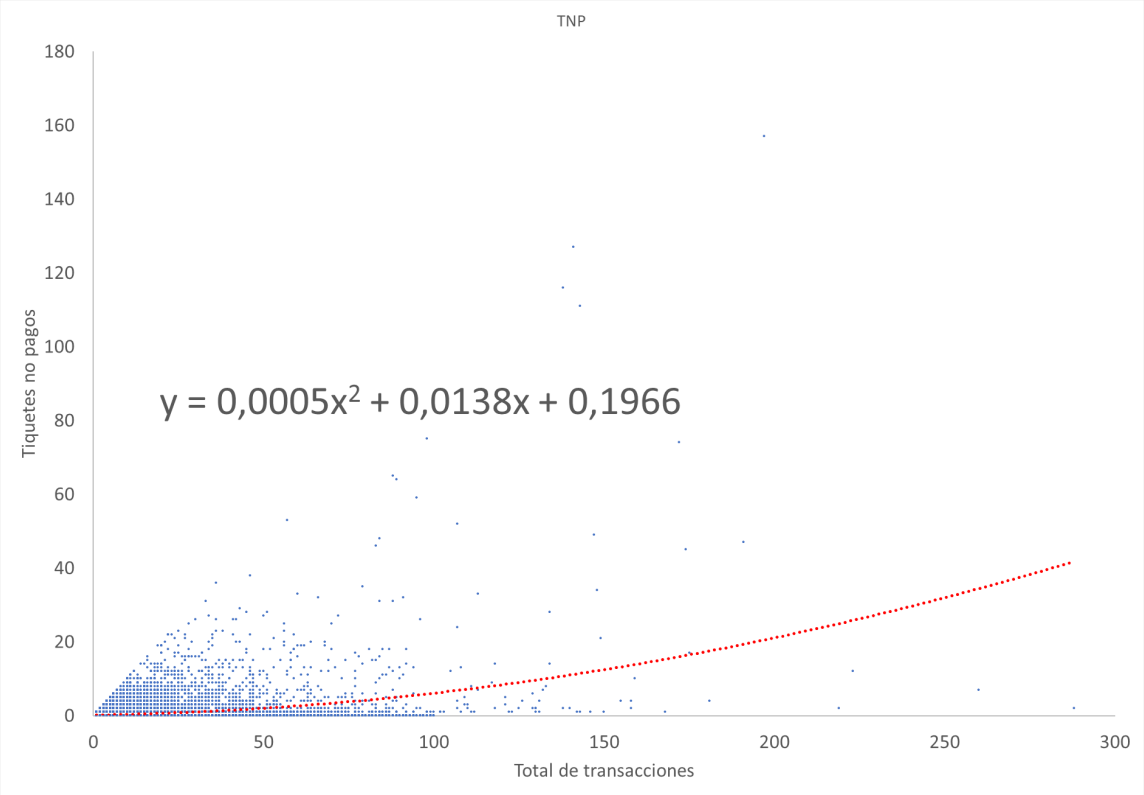

Figura 7. Caracterización del score por placa para fraude.

Para la definición del *score* se tuvo en cuenta el comportamiento de las transacciones no pagas dentro del total de transacciones totales de cada uno de los clientes. Teniendo en cuenta esto, en la Figura 7 se muestra la relación entre las Transacciones No pagas respecto Transacciones totales de cada cliente. Se elige hacer una regresión polinomial de grado dos ya que se quería distinguir en el *score* aquellos clientes con pocas transacciones respecto a los clientes regulares. La lógica aplicada, es que todos los puntos arriba de la curva deben tener un *score* negativo (implicando un mal comportamiento en el hábito de pago) y viceversa. Para lograr lo anterior, se define el score como

$$score_{fraude} = 0.0005Tx^2 + 0.0138Tx + 0.1966 - TxNP$$

donde

$$\begin{cases} Tx : Total\ de\ transacciones \\ TxNP : Transacciones\ no\ pagas \end{cases}$$

Adicionalmente, se realizó una limpieza en términos de negocio a aquellos clientes que no tengan suficiente historia para tener una calificación adecuada. Así, no penalizar a clientes que puedan tener pocas transacciones y alguna no paga. Por lo tanto, por términos del negocio se eliminaron los clientes que no cumplan con el mínimo de 5 transacciones. En esta reducción de clientes se pasa de tener 98967 placas a 20452.

En la Figura 8 se presenta el diagrama de flujo de los datos para ser preparados, explorados y analizados.

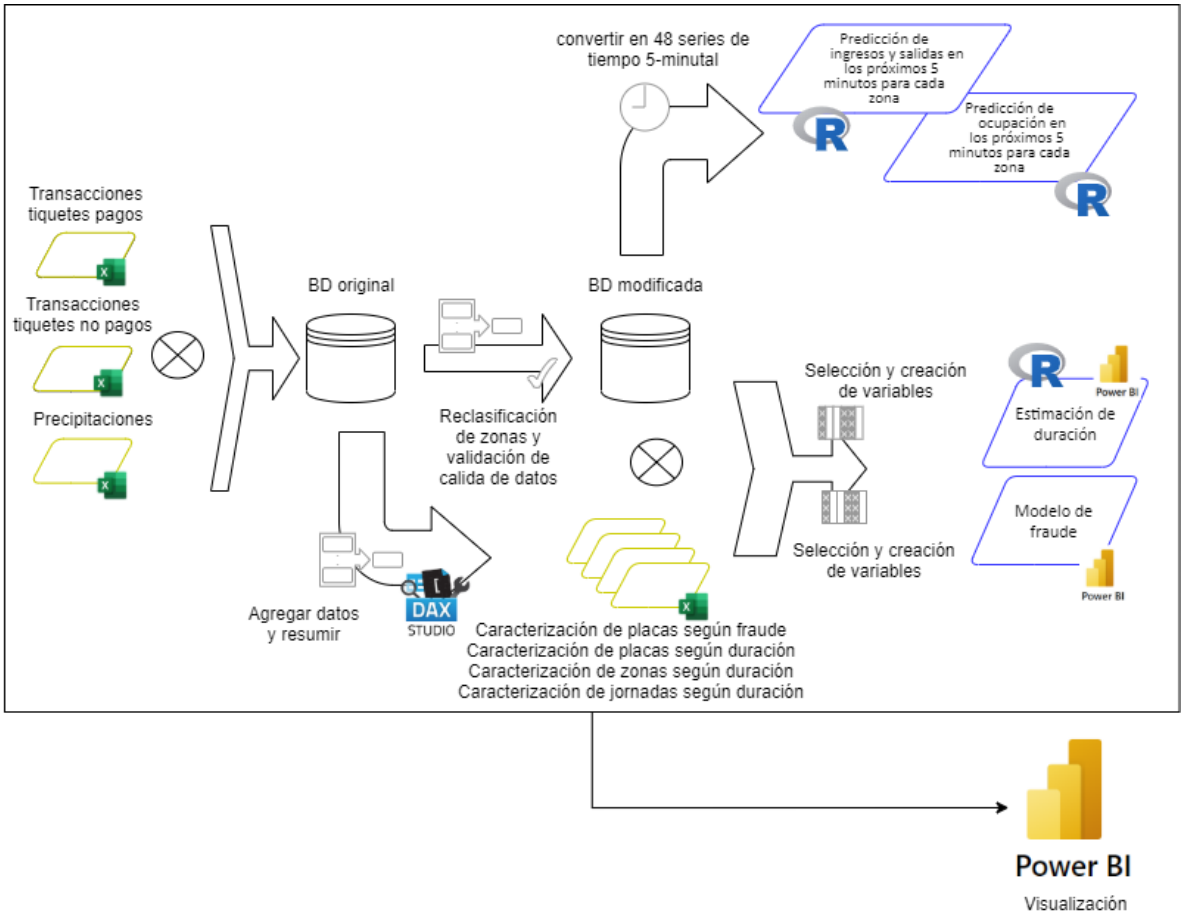

Figura 8. Diagrama de flujo de preparación, manipulación y exploración de datos. Elaboración propia.

4. Modelamiento

Como se mencionó al final de la sección 1 del presente documento, se utilizarán modelos de series de tiempo, modelos de aprendizaje supervisado y modelos de clasificación binaria para abordar cada uno de los objetivos de analítica del proyecto. En esta sección se detallan los modelos desarrollados, los protocolos de experimentación y evaluación de cada modelo.

Para aclarar, en la arquitectura presentada en la Figura 8, se observa una base de datos inicial llamada “Precipitaciones”. El propósito de esta base de datos era utilizar datos climáticos como regresores exógenos a las transacciones. Sin embargo, no fue posible utilizar datos cercanos al área

geográfica donde se ubican las ZER. El lugar más cercano encontrado fue el Aeropuerto Olaya Herrera en Medellín, ubicado a 22 km del municipio de Caldas, Antioquia. Dada la lejanía geográfica de los datos, se observó en los modelos que esta variable de precipitación no tenía efecto sobre las variables objetivo según el modelo. Lo anterior no quiere decir que la precipitación sobre Caldas no tenga relevancia para lograr los objetivos de analítica planteados, sino que los datos disponibles al público que fueron utilizados no tienen relación con las condiciones reales de Caldas.

4.1. Modelos de series de tiempo para predicción de ingresos y salidas de los próximos 5 minutos

Inicialmente, hay que tener presente los sucesos del último año, cómo la situación de la pandemia cambió las dinámicas sociales y organizacionales en todo el mundo. En la Figura 5 se muestra una vista general de cómo cambió la dinámica en cuanto a utilización de ZER en el municipio de Caldas en Antioquía. Se ve que el uso fue a cero durante el periodo de cuarentena general en Colombia, en este caso del 24 de marzo al 15 de abril del 2020. Posterior a esta fecha sigue un periodo de reactivación económica en donde el uso de ZER creció, y hasta agosto del 2020 se puede decir que alcanzó el punto estable. Dado esto, las series de tiempo desarrolladas en este proyecto utilizarán datos desde el 1 de agosto hasta el 16 de octubre del 2020 y se utilizarán como datos de prueba los últimos 15 días de octubre del 2020.

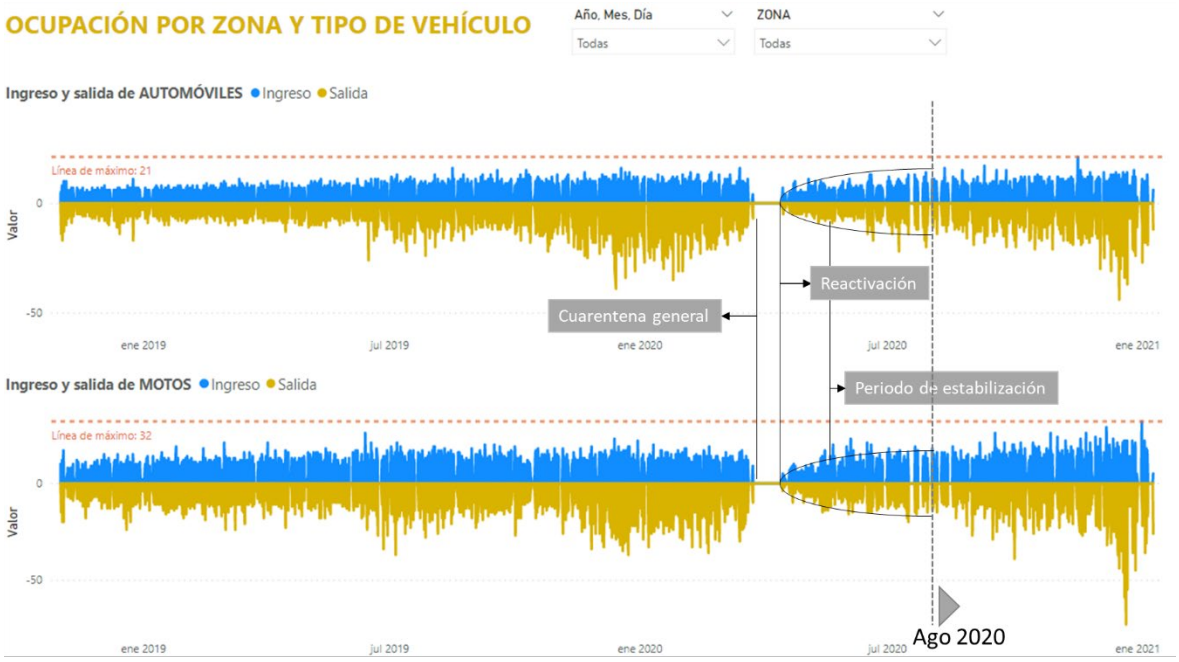

Figura 9. Vista general de la serie de tiempo de ingreso y salida 5-minutal.

Para el modelado de serie de tiempo, es mandatorio establecer las características principales de ésta, características tanto de construcción de la serie, como visuales. En la Figura 10 se presenta un ejemplo del tipo de serie de tiempo que se modela para, posteriormente, realizar los pronósticos. Este ejemplo en particular es la serie de tiempo de ingreso de motos de la zona PARQUE CALDAS, con una ventana de dos meses, desde septiembre a octubre del 2020. De la Figura 10 se destacan dos aspectos: la serie de tiempo tiene una temporalidad de 15 días, que representa la dinámica

social de las quincenas de pago, y que la serie de tiempo tiene gran cantidad de ceros. De hecho, estas series de tiempo son de conteo. Los profesores Rob J. Hyndman y George Athanasopoulos en el libro *Forecasting: Principles and Practice* [1] explican por qué este tipo de serie de tiempo debe abordarse diferente a las convencionales.

Los conteos son números enteros y, por lo general, de poca magnitud, en el caso del ejemplo de la Figura 10, el máximo valor es 20. Lo anterior quiere decir que el fenómeno no está en un espacio continúo, sino en uno discreto y al ser de baja magnitud la solución de redondear al entero más cercano puede agregar error considerable al modelo y al pronóstico [1]. Sin Embargo, los modelos de serie de tiempo de conteo, tales como *Croston* o *Zero-Inflated Models*, pronostican la cantidad de demanda (ingresos o salidas en este caso) acumulada que habrá en un tiempo específico [1]. Por lo anterior este tipo de modelos no lograría satisfacer el objetivo de analítica planteado.

Por último, una temporalidad de 15 días, en una serie de tiempo 5-minutal, implica que por temporada hay 4320 datos (cantidad de ventanas de 5 minutos contenidos en 15 días). Esto causa un problema a la hora de modelar la serie de tiempo con ARIMA o ETS ya que estos modelos no soportan temporalidades mayores a 350 datos [2], que está muy por debajo de la serie de tiempo en cuestión. Para disminuir la cantidad de datos en cada temporada, se sabe que el servicio de ZER inicia a las 07:00 horas y finaliza a las 19:00 horas, por lo que es viable parametrizar el día de la serie de tiempo entre estas horas. Lo anterior hace que la cantidad de datos disminuya al 50%, es decir 2160, que sigue siendo gran cantidad para ser soportada por ARIMA o ETS.

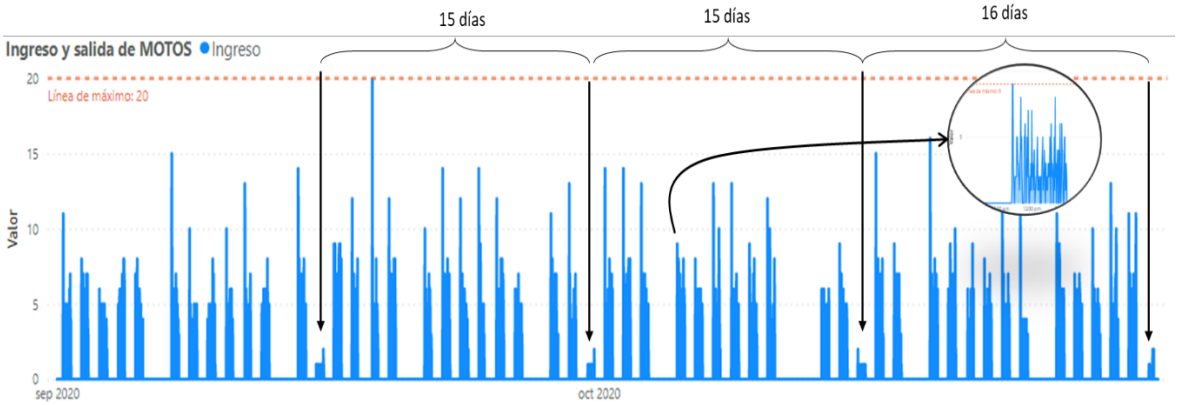

Figura 10. Ejemplo de características de serie de tiempo.

Explicadas las características de estas series de tiempo y las restricciones que tienen, a continuación, se presentan lo modelos de serie de tiempo desarrollados, cómo se realizó el protocolo de experimentación y la evaluación de cada uno. El paquete *forecast* de R permite realizar una descomposición temporal y de tendencia y realizar pronósticos sobre esta descomposición utilizando la función *stlf()*. La función hace un pronóstico sobre el componente de temporalidad ajustado utilizando un método definido por el usuario: ETS, ARIMA, naive y random walk [1]; de esta manera se obtuvo 4 modelos distintos. Por otro lado, también está la función *snaive()*, que hace un pronóstico ingenuo tomando el mismo valor de la temporada anterior. Adicionalmente, se implementó el procedimiento *prophet()* del paquete del mismo nombre. Prophet es un desarrollo de Facebook para el pronóstico de series de tiempo dando la posibilidad al desarrollador de agregar regresores exógenos a la serie y definir temporalidades según la necesidad [3]. El último modelo

realizado se basó en estadística descriptiva. Tanto para el modelo de estadística descriptiva como el de prophet, se utilizaron las siguientes variables exógenas:

- Día de la semana: de domingo a sábado
- Hora del día: Ventana 5-minutal de 07:00 horas a 19:00 horas.
- Tipo de día: puede ser día festivo, día después de festivo o día de pago
- Temporalidad diaria (sólo en el modelo prophet)
- Temporalidad semanal (sólo en el modelo prophet)
- Temporalidad quincenal (sólo en el modelo prophet)

El pronóstico del modelo de estadística descriptiva toma como pronóstico el promedio de los últimos 4 meses (datos de entrenamiento) discriminando por cada uno de los factores descritos arriba. Por último, la serie de tiempo se construyó en R con la función *ts()* con una frecuencia de 2016 datos (7 días). Se define la temporada de 7 días pues la temporada semanal es más influyente que la temporada quincenal expuesta en la Figura 10.

Como se explicó más atrás, la base de entrenamiento utiliza datos del 1 de agosto al 16 de octubre del 2020, y la de pruebas será del 17 de octubre al 31 de octubre del mismo año. Para cada base se calculan las métricas de MAE, MASE, ME y RMSE bajo los mismos parámetros expuestos en esta sección. En la Figura 11 se presentan las métricas de desempeño de los 7 modelos desarrollados. Dado que el MASE no tiene escala, sirve para comparar distintos pronósticos de la misma serie de tiempo y también se puede promediar los MASE de pronósticos de diferentes series de tiempo para caracterizar diferentes modelos [4]. Por lo anterior, la métrica MASE se ajusta a la necesidad del proyecto pues, como se ha explicado, se evaluará cada modelo sobre 48 series de tiempo distintas. Las métricas en base de entrenamiento se presentan para mostrar que no hay sobreajuste en los modelos.

Métricas de desempeño generales (promedio de los resultados de las 48 series de tiempo analisadas)

train/test	SNAIVE	STL_ARIMA	STL_ETS	STL_NAIVE	STL_RW	STAT	PROPHET
train							
RMSE	0,77	0,50	0,51	0,69	0,69	0,52	0,56
ME	0,00	0,00	0,00	0,00	0,00	0,04	0,06
MASE	1,00	0,94	0,96	1,18	1,18	0,60	0,62
MAE	0,37	0,30	0,30	0,37	0,37	0,23	0,25
test							
RMSE	0,82	0,68	0,68	0,74	0,75	0,58	0,66
ME	0,08	0,00	-0,01	0,12	0,13	0,08	0,10
MASE	1,12	1,08	1,08	0,96	1,00	0,60	0,66
MAE	0,39	0,36	0,36	0,40	0,41	0,27	0,31

Figura 11. Métricas de desempeño para los modelos de pronóstico de ingresos y salidas en los próximos 5 minutos.

Se encuentra que el modelo basado en estadística descriptiva, con información exógena como lo es el tipo de día, obtiene las mejores métricas de desempeño. Con un MASE de 0,60 en la base de prueba demuestra ser mejor que el pronóstico ingenuo y, además, muestra tener menor error medio cuadrático de 0,52, es decir, más o menos una persona que entra o sale de la ZER. Cómo se

mencionó previamente, comparando las métricas de entrenamiento y prueba, se puede concluir que los modelos no presentan sobreajuste.

4.2. Modelos de estimación de duración

Para el modelo de regresión se utilizó el paquete de R *glmnet* para implementar regresiones lineales. Con este paquete se desarrollaron 3 modelos: step-wise, regresión con regularización tipo Lasso y regresión con regularización tipo Ridge. Adicionalmente, utilizando el paquete *e1071* de R para implementar un modelo de Vectores se Soporte Regresores. Por último, similar al modelo de series de tiempo, se hizo un modelo de estadística descriptiva que obedece a cada elemento del cubo de datos mostrado en la Figura 6, es decir, la regresión se hacía evaluando la placa, la zona y la jornada de la transacción y se buscaba esa posición en el cubo.

Para el desarrollo de los modelos, se utilizaron variables obtenidas del cubo de información mostrado en la Figura 6, además de otras variables sintéticas o disponibles en la base de datos transaccional. Específicamente, de la base de datos transaccional se utilizó la hora de ingreso del vehículo y el día de la semana de la transacción; del cubo de información se tomó la duración promedio de la placa de la transacción y la duración promedio de la zona y jornada correspondiente y; por último, se utilizó el *score* de fraude construido y explicado en la sección 3.3 del presente documento.

Antes de presentar los resultados, en la Figura 12 se muestra la distribución de la variable duración. Se observa que presenta una simetría positiva por lo que se recomendaría aplicar una normalización antes de empezar al modelo. También se muestra la misma variable duración aplicándole la normalización logarítmica, pareciendo una distribución más normal que la variable original. Sin embargo, este caso particular, al hacer el modelo con la variable normalizada y evaluar el error de cada modelo, el error aumenta en lugar de disminuir respecto a los modelos con la variable original. Por lo anterior, los resultados presentados a continuación corresponden a modelos que estiman directamente la duración, sin transformación logarítmica.

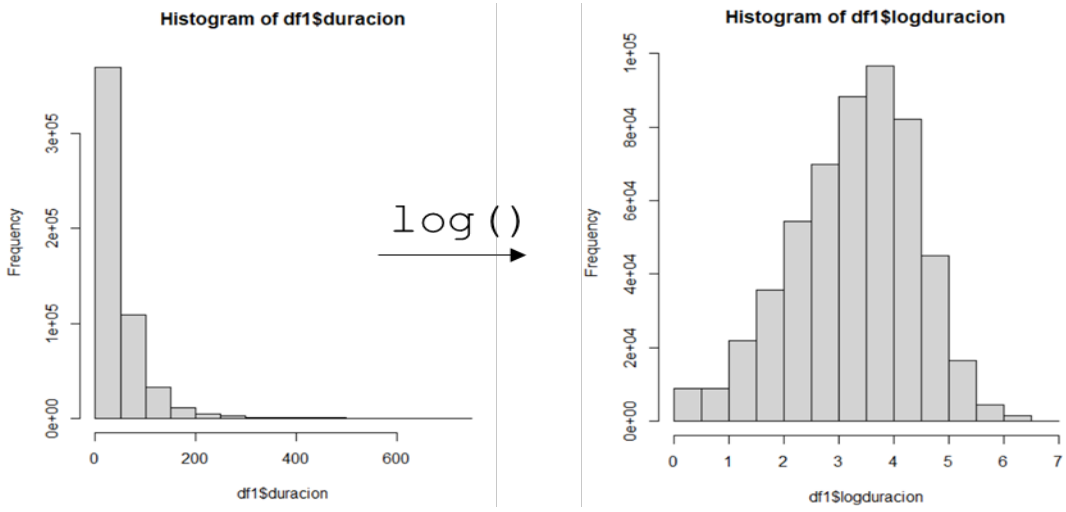

Figura 12. Histograma de variable original y normalizada.

Para el entrenamiento de los modelos se utilizó el 80% para la base de entrenamiento y el 20% como prueba. La métrica seleccionada para la comparación de modelos es el RMSE, ya que es una métrica

que está en unidades de la medida, es decir, en minutos. En la Tabla 2 se presentan los resultados obtenidos.

Tabla 2. RMSE medido en minutos para cada modelo.

Modelo	**RMSE (minutos)**
Step-wise	41,51
Regularización Lasso	41,51
Regularización Ridge	41,54
Vectores de soporte regresores	69,64
Estadística descriptiva	32,39

Se encuentra, nuevamente, que el modelo que presenta menor error es el estadístico descriptivo, con un error de 32 minutos. En términos de negocio, un modelo que representa valor agregado es aquel que logre un error menor a 5 minutos. Por lo tanto, los modelos encontrados no representan valor agregado a la organización.

4.3. Modelos de Fraude

Luego de la limpieza de datos realizada y de crear la variable de *score,* se procedió a usar la visualización de elementos influyentes claves de Power BI. Esta herramienta implementa una regresión logística para obtener la importancia de cada una de las variables predictoras. Con el coeficiente obtiene los elementos más influyentes. Después, sobre los tres elementos más influyentes empieza a particionar la base de datos hasta encontrar los segmentos que mejor dividan la variable a predecir, en este caso, si la transacción fue paga o no [5].

Se agregó al objeto visual de Power BI, además del *score*, variables como Tipo de Vehículo, Hora de ingreso, Total Transacciones, Zonas y Duración para poder identificar el conjunto de variables que dan más probabilidad de que la transacción no vaya a ser paga. En la Figura 13 se presenta la visualización obtenida de elementos influyentes clave.

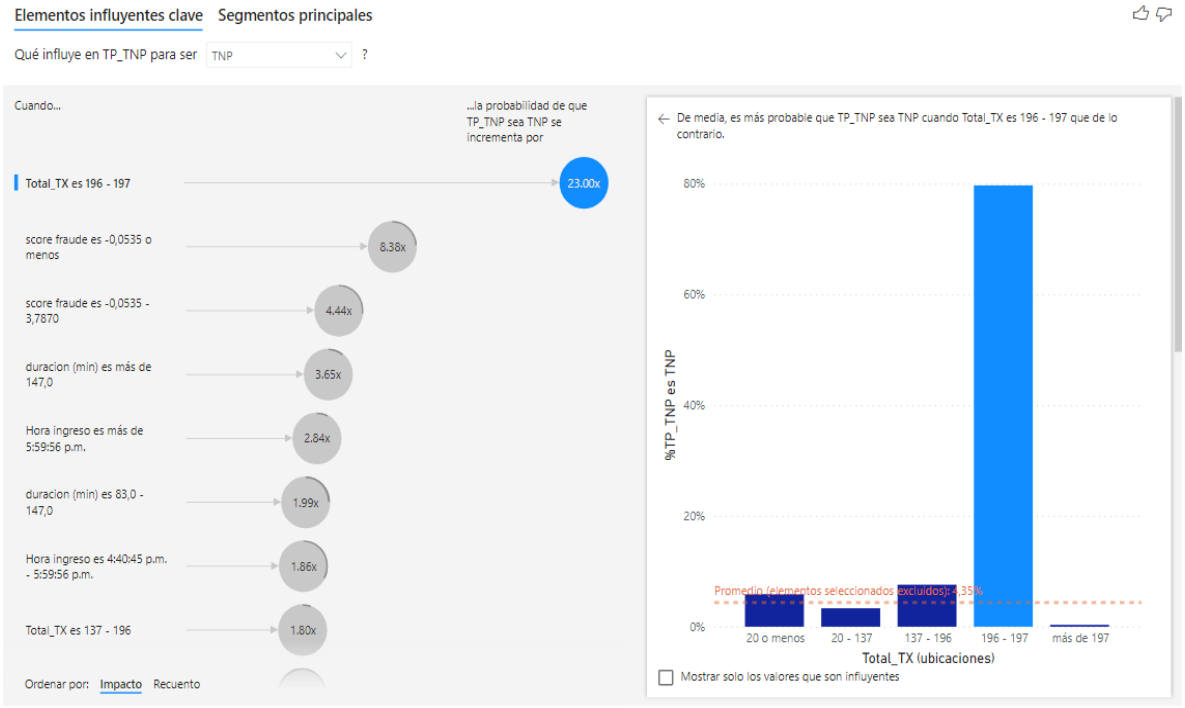

Figura 13. Elementos influyentes calve para ser transacción no paga.

Por ejemplo, lo que presenta la visualización en el primer caso es que una transacción que corresponda a un cliente en el que el *score* sea menor a –0,0535 tendrá 8,36 más de probabilidad de que la transacción sea no paga, en la segunda línea muestra que en caso de que las transacciones totales de ese cliente sean entre 196 y 197 tiene 23 veces más de probabilidad de que la transacción sea no paga (en comparación cuando no cumple esa condición), y así con cada uno de los elementos claves encontrados.

Por otro lado, en la Figura 14 se presentan la cantidad de segmentos principales encontrados según la partición de los elementos más influyentes y que cantidad de elementos en el segmento son transacciones no pagas. Por ejemplo, el segmento 1 cuenta con 58,2% de transacciones no pagas de un total de 4233.

Cabe destacar que para este modelo se tuvo en cuenta la duración del vehículo dentro de la ZER ya que esta variable genera una mejor caracterización de los segmentos principales, sin embargo, sabemos que es imposible obtener esta variable hasta que la transacción haya terminado.

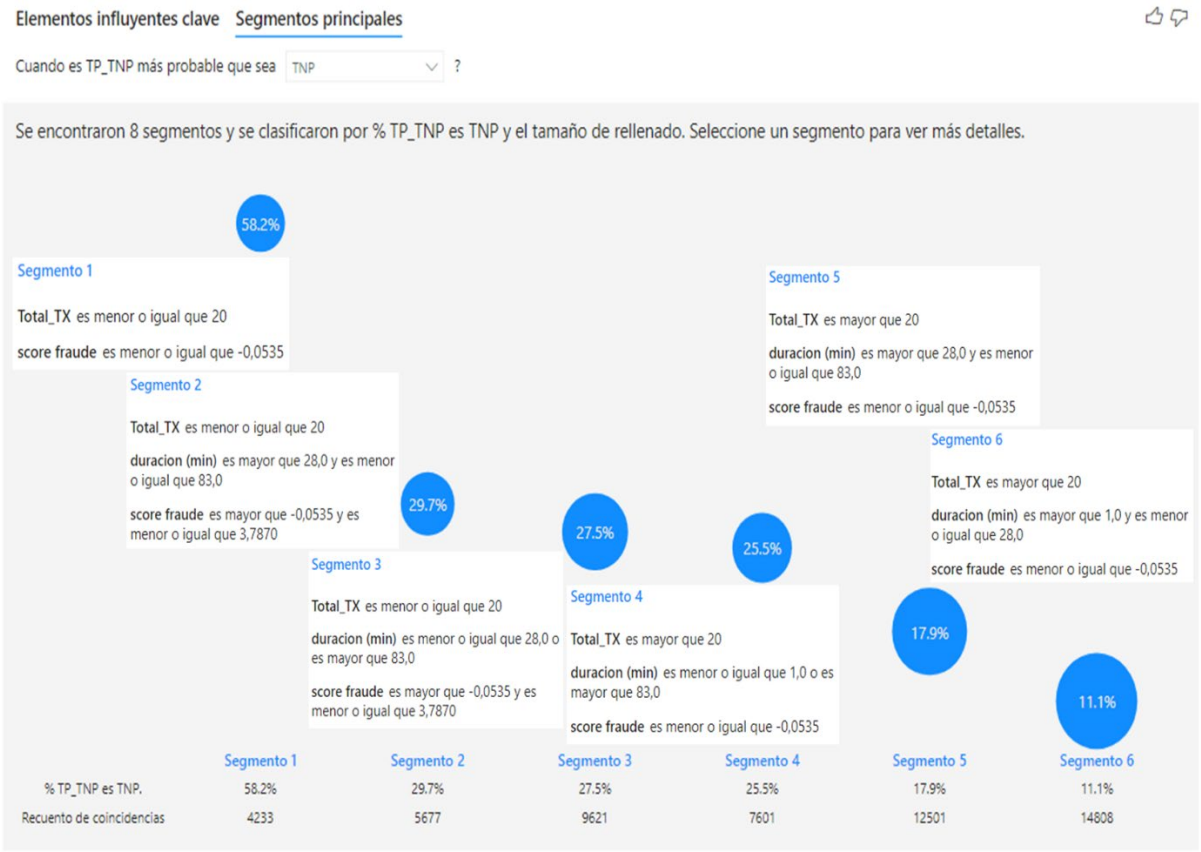

Figura 14. Segmentos principales para que una transacción sea no paga.

Ahora, para entender los segmentos principales encontrados, en la Figura 15 se muestra el primer segmento principal. Las características de este segmento son: El total de transacciones tiene que ser menor o igual a 20, el *score* sea menor o igual a –0,0535. La información que da este segmento es que, si una transacción contiene estas características, tendrá una probabilidad de 58,2% de que la transacción sea no paga.

Adicionalmente, es posible encontrar más información que pueda ser relevante para el negocio ya que brinda algunas ayudas para la toma de decisiones. Lo que se puede visualizar en las Figura 16 y Figura 17, es si existe alguna diferencia con respecto con otras variables, en este caso las zonas. La manera correcta de interpretar estos resultados es si la transacción del segmento 1 se da en PARQUE CALDAS hay una mayor probabilidad de que la transacción sea paga, en cambio, si la zona es FULLHOGAR el comportamiento es contrario, ya que al usar esta zona tendrá mayor probabilidad de no pago de la transacción.

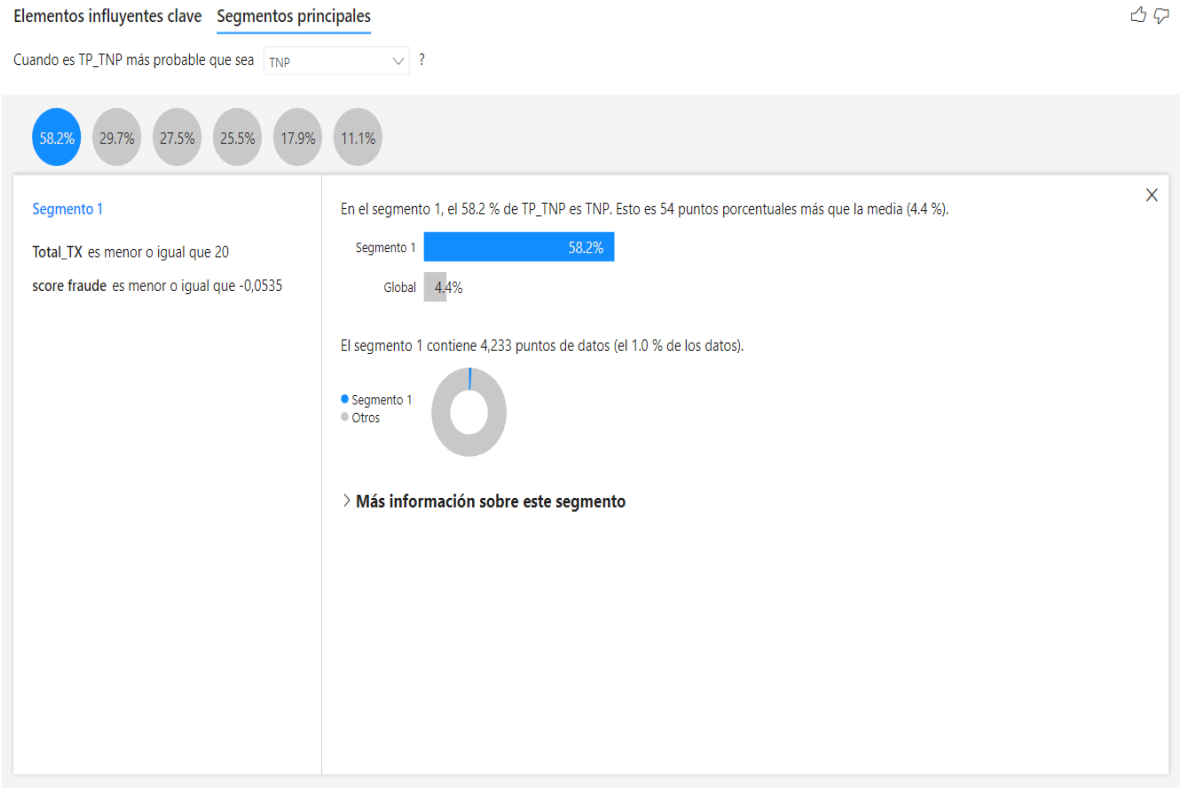

Figura 15. Características del segmento 1.

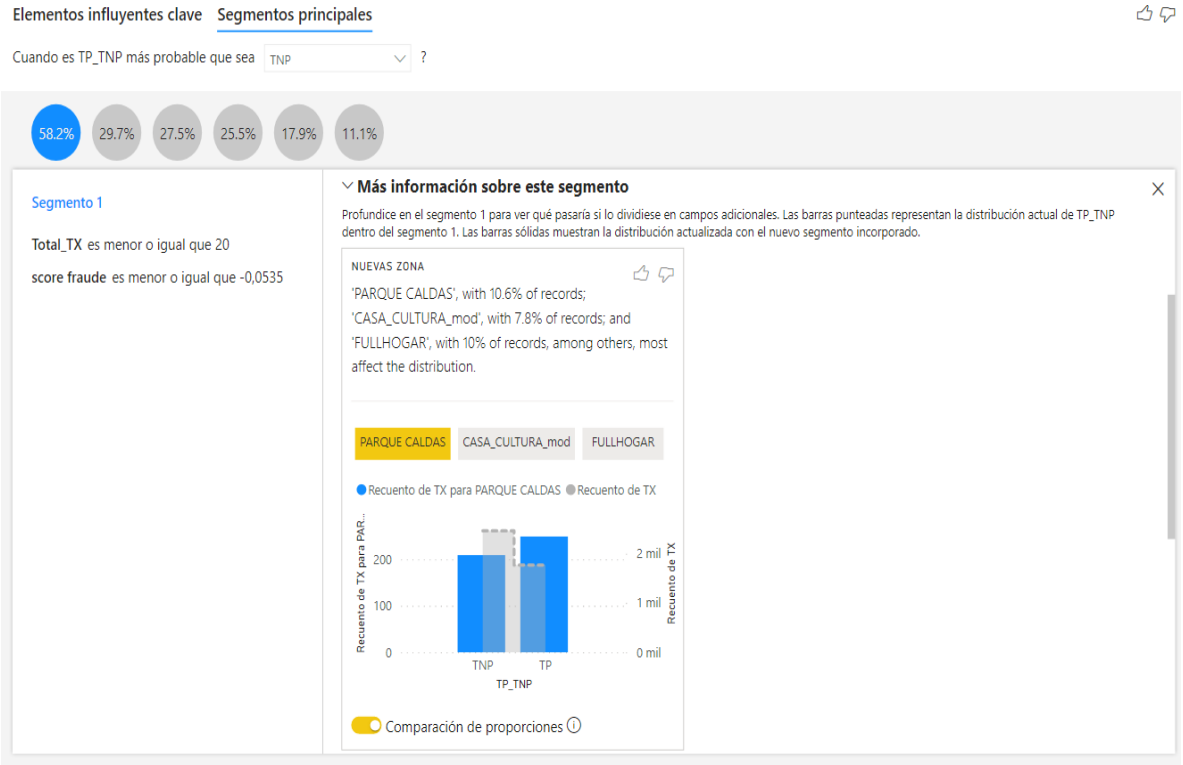

Figura 16. Características del segmento 1, diferenciado por zona PARQUE CALDAS.

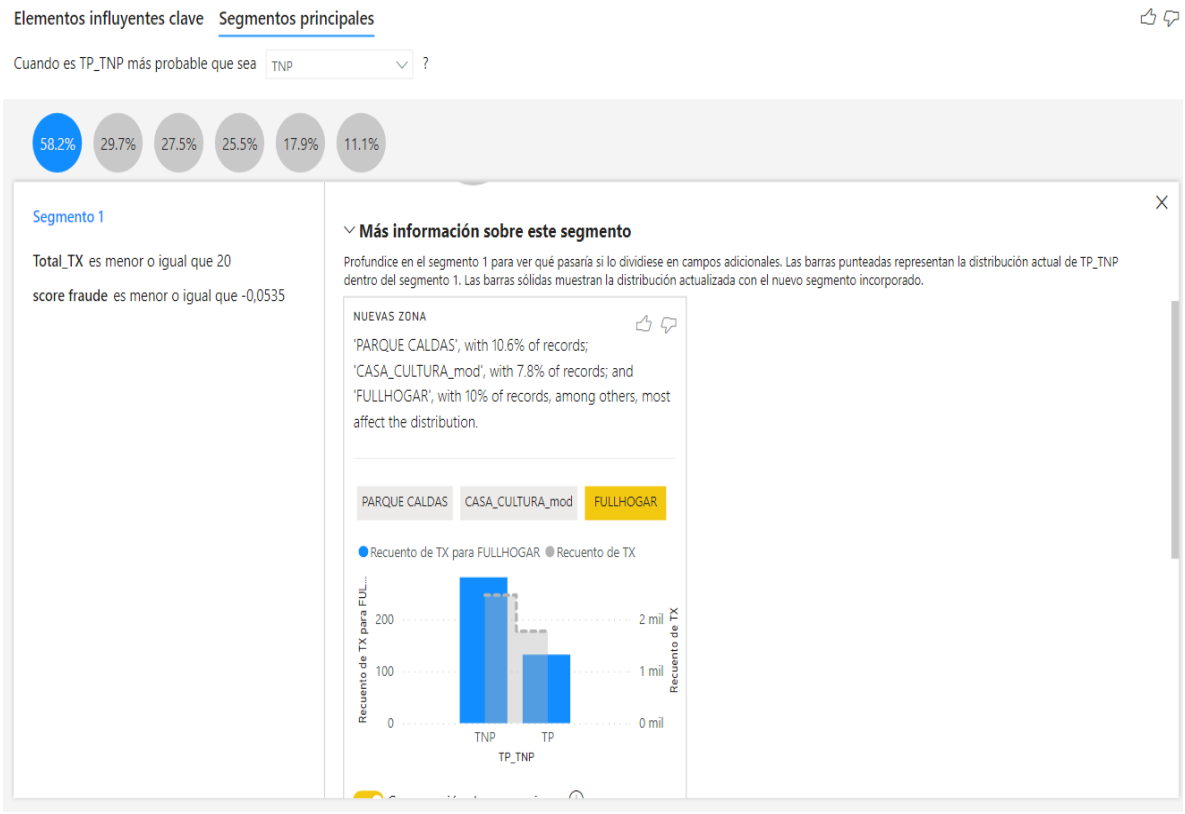

Figura 17. Características del segmento 1, diferenciado por zona FULLHOGAR

Ahora se procederá a resumir la información para cada uno de los segmentos restantes:

El **Segmento 2:** Aquellas transacciones cuya placa tiene un total de transacciones menor o igual que 20, la duración sea mayor que 28 minutos y menor que 83 minutos y que el *score* sea mayor que -0,535 y menor o igual que 3,7870. La información que da este segmento es que, si una transacción contiene estas características, tendrá una probabilidad de 29,7% de que la transacción sea no paga.

El **Segmento 3:** Aquellas transacciones cuya placa tiene un total de transacciones menor o igual 20, la duración menor o igual que 28 minutos o mayor que 83 minutos y el score mayor que –0,535 y menor o igual que 3,7870, si una transacción cumple estas características tendrá una probabilidad de 27,5% de que la transacción sea no paga.

El **Segmento 4:** Aquellas transacciones cuya placa tiene un total de transacciones mayor que 20, la duración menor o igual que 1 minuto o mayor que 83 minutos y el *score* menor o igual que –0,535. Una transacción que cumpla esas características tendrá una probabilidad de 25,5% de que la transacción sea no paga.

El **Segmento 5:** Aquellas transacciones cuya placa tiene un total de transacciones mayor que 20, la duración mayor que 28 minutos y menor o igual que 83 minutos y el *score* menor o igual –0,0535. Una transacción que cumpla esas características tendrá una probabilidad de 17,9% de que la transacción sea no paga.

El **Segmento 6:** Aquellas transacciones cuya placa tiene un total de transacciones mayor que 20, la duración tiene que ser mayor que 1 minuto y menor o igual que 28 minutos y el *score* tiene que ser menor o igual que –0,0535, este segmento es el que menor da probabilidad de no pago, ya que tiene una probabilidad de 11,1%.

En resumen, cada uno de los segmentos identificados tiene algunas características principales que explican que si la transacción cumple esas características tendrá una probabilidad asociada a el no pago. Adicional a esto, dentro de cada segmento se tiene otras variables que pueden modificar la probabilidad, tales como la zona de la transacción o tipo de vehículo.

5. Evaluación

El modelo recomendado para abordar el pronóstico 5-minutal de ingreso y salida de las ZER es el basado en estadística descriptiva. Dado que se trata de estadística descriptiva, este es un modelo de baja complejidad computacional y que puede ser parametrizado a medida que se encuentren más factures influyentes en el comportamiento de las series de tiempo. Como se presentó en la Figura 11 el error medio cuadrático es aproximadamente de 1, lo que quiere decir, que en promedio se sabrá con una certeza de más o menos una persona que en determinada zona ingresarán o saldrán $x \pm 1$ carros o motos. Con esta herramienta, el cliente estará en capacidad de optimizar las rutas de los promotores de manera que estén en la zona cuando estén llegando o saliendo vehículos, ya sea para iniciar la transacción o para finalizarla. Por último, se recomiendo al cliente recolectar información hidrológica que corresponda a la zona geográfica de operación del negocio para ser utilizada como regresor exógeno en los modelos.

Ahora, para el modelo de duración, con los datos disponibles y las variables sintéticas creadas no se encontró un modelo que agregara valor al negocio. Con un error de 32 minutos no sería viable tomar decisiones de negocio basado en los modelos acá presentados. Por lo anterior, se recomienda al cliente iniciar a recolectar diferentes variables que puedan aportar a estimar la duración de un vehículo parqueado en una ZER. En particular, se recomienda recolectar información de marca del vehículo y estado físico visual del vehículo para aprovecharlas como un indicador de capacidad monetaria. El estado físico del vehículo se podría establecer con categorías combinando estados binarios de golpes y rayones. Otra recomendación es agregar en la recolección de datos el género del conductor complementando con una variable los tipos de comercios aledaños en las zonas para ayudar a los modelos a identificar duración en estos establecimientos.

Por último, para el modelo de fraude hay que tener en cuenta que es un modelo exploratorio y no un modelo predictivo ya que al no tener variables que pudieran explicar algún tipo de comportamiento del cliente tales como tipo de carro, modelo de carro o específicamente de la persona, se tuvo que explicar el comportamiento del cliente mediante las transacciones históricas, como se presentó en el modelo y la explicación de lectura del Power BI. Este modelo es muy útil para la explicación del negocio ya que permite a Generación Móvil hacer uso eficiente de los promotores al tener ya una base de datos del Score del cliente y además variables predictoras como las zonas, implementar alertas de clientes que por historial tienen un comportamiento de no pago o algunas zonas en las cuales su decisión de no pago aumenta. Por lo tanto, ayudaría a ubicar de manera analítica sus promotores para reducir la cantidad de transacciones no pagas reduciendo así pérdidas monetarias, también es importante resaltar que las variables sintéticas son históricas y que

vienen de las transacciones realizadas por cada cliente durante la historia del negocio, por lo tanto, a mayor historia se genera una mejor caracterización de las placas y por lo tanto del Score calculado, generando que el Score del cliente sea dinámico a través del tiempo.

6. Bibliografía

[1] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*, 2nd ed. Melbourne, Australia: OTexts, 2018.

[2] R. J. Hyndman, "Forecasting with long seasonal periods," 2010. https://robjhyndman.com/hyndsight/longseasonality/ (accessed Apr. 11, 2020).

[3] Facebook Open Source, "Prophet. Forecasting at scale." https://facebook.github.io/prophet/ (accessed Apr. 15, 2021).

[4] R. J. Hyndman, "Another look at forecast-accuracy metrics for intermittent demand," *Int. J. Energy Stat.*, vol. 04, no. 02, pp. 43–46, 2016, doi: 10.1142/s2335680416500083.

[5] Microsoft, "Create key influencers visualizations," 2021. https://docs.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-influencers.