Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

Factores determinantes de la tenencia de automóviles en Argentina

Un análisis multivariado de los registros del automotor para la gestión eficiente de datos en la Asociación de Fábricas Automotores

AUTORA: KARINA BARTOLOMÉ
MENTORA: NATALIA SALABERRY

DICIEMBRE 2021

Resumen

El mercado automotor suele estudiarse desde dos enfoques: oferta y demanda. En este trabajo se analizará el enfoque de demanda, haciendo énfasis en la naturaleza de bienes duraderos de los vehículos. Se busca entender el tiempo que transcurre hasta que un individuo vuelva a ser demandante de este tipo de bien. En este sentido, se evalúan las características que influyen sobre la duración entre la inscripción y la baja de un automotor. Dada la distribución bimodal de la duración de los vehículos, se realiza una clasificación en corta y larga duración. Se utilizan datos de bajas en registros seccionales en Argentina entre 2018 y 2021, así como también, datos adicionales que aporten valor al caso de análisis. Luego de caracterizar la situación, se realiza una estimación del efecto de diversos factores sobre la clasificación establecida. Para ello, se utilizan modelos estadísticos multivariados y de aprendizaje automático. Considerando los resultados obtenidos, se establece un método para la gestión de datos de nuevas inscripciones, con el objetivo de mejorar la toma de decisiones de la Asociación de Fábricas Automotores.

Palabras clave: Tenencia automovilística, Asociación de Fábricas Automotores, Registros seccionales, Métodos multivariados, Gestión eficiente

Índice

Introducción 3
1. La gestión de datos en el sector automotor argentino 5
1.1. El rol de las organizaciones en la tenencia automovilística: el caso de ADEFA y la DNRPA 5
1.2. El impacto de los grandes volúmenes de datos en las organizaciones 6
1.3. Situación de la demanda de vehículos para uso personal 8
2. Métodos predictivos aplicados a la duración de la tenencia de automóviles 12
2.1. Metodología para clasificar a los vehículos en corta y larga duración 12
2.2. Análisis exploratorio y procesamiento de datos 17
2.2.1. Automotores 17
2.2.2. Titular 21
2.2.3. Contexto 22
2.3. Determinación de los factores que explican la duración de la tenencia de vehículos en Argentina 23
3. Implementación de modelos para la toma de decisiones 30
3.1. Aplicación de modelos analíticos predictivos 30
3.2. Análisis de resultados de las predicciones obtenidas 31
3.3. Recomendaciones para una gestión eficiente de datos 35
Conclusión 37
Referencias bibliográficas 39
Apéndice A: Organizaciones 41
Apéndice B: Estado de situación del mercado automotor argentino 43
Apéndice C: Modelo de clasificación de la duración 45
Reporte de la mentoría 51

Introducción

El mercado automotor argentino cuenta con gran cantidad de actores involucrados, dada la importancia del sector en términos de generación de empleo, producción y participación en la movilidad de los ciudadanos. En este contexto, se destaca el rol de dos grandes organizaciones: la Dirección Nacional de los Registros Nacionales de la Propiedad del Automotor (DNRPA)[1] y la Asociación de Fábricas de Automotores (ADEFA)[2]. La DNRPA es la entidad encargada de la regulación de la propiedad automotor, administrando gran cantidad de datos vinculados a la tenencia de vehículos: inscripciones, transferencias, denuncias de robos y bajas. Por otro lado, la ADEFA agrupa a doce empresas que producen automóviles en Argentina, fomentando la generación de empleo y la mejora en la competitividad de estas organizaciones.

Para realizar estudios del mercado automotor se utilizan dos perspectivas: la oferta y la demanda. Dado que la industria automotriz es un sector fundamental en la estructura productiva argentina, existe amplia literatura que estudia el comportamiento de la oferta. Por otro lado, la literatura vinculada a la demanda no es tan amplia. Los automóviles se caracterizan por ser bienes duraderos, es decir, una vez adquiridos pueden ser utilizados un gran número de veces a lo largo del tiempo. En este sentido, para entender las preferencias en este mercado es necesario analizar el tiempo que transcurre desde la inscripción inicial de un vehículo hasta el momento en que es dado de baja.

El objetivo de este trabajo es analizar la dinámica de la demanda de automóviles en Argentina. De esta forma, se busca responder al siguiente interrogante: ¿cuáles son los factores que explican el tiempo de tenencia? Particularmente, se analizará el caso específico de la duración de los vehículos producidos por las organizaciones agrupadas en la ADEFA. Para ello, se utilizarán datos de inscripciones y bajas en registros seccionales, así como también, datos adicionales que aporten valor al caso de análisis. Al analizar las bajas registradas entre los años 2018 y 2021, se observa una distribución bimodal en la duración de la tenencia, con lo cual se clasifican las observaciones en vehículos de corta o larga duración. En este sentido, se utilizarán modelos estadísticos multivariados y de aprendizaje automático para clasificar las nuevas inscripciones. Se hará énfasis en los modelos de vehículos producidos por empresas nacionales en el año 2021, es decir, automotores relevantes para la ADEFA. De esta forma, se propondrá

[1] Acerca de DNRPA: https://www.dnrpa.gov.ar/portal_dnrpa/
[2] Acerca de ADEFA: http://adefa.org.ar/

un método para estimar la cantidad de vehículos inscriptos en el año 2021 que tendrán una baja anticipada, desagregando los resultados en los distintos modelos de automotores. Las bajas anticipadas registradas son en gran parte con recuperación de autopartes. Por esta razón, es de gran relevancia para las organizaciones del sector autopartista estimar cuáles son los automotores que requerirán una mayor producción de autopartes. De esta forma, los resultados serán de utilidad para las organizaciones involucradas, permitiendo mejorar la toma de decisiones en base a evidencia.

El trabajo está organizado de la siguiente forma. En primer lugar, se realizará un análisis del estado de situación del mercado automotor argentino. Para ello, se presentarán las organizaciones involucradas en este sector, junto con las diversas fuentes de datos a contemplar. Se mencionarán aspectos vinculados a la gestión de datos en este sector, junto con un análisis del contexto en el cual se enmarca la problemática a trabajar, que será detallada junto con aspectos metodológicos a considerar. En el caso particular de los automotores, se aplicarán técnicas estadísticas para obtener variables a partir de los modelos de vehículos: análisis de componentes principales y *clustering*. Luego de generar una base de datos adecuada para la resolución de la problemática, se ajustarán modelos estadísticos multivariados y de aprendizaje automático para clasificar a los vehículos en dos categorías según su duración. Se seleccionará el mejor modelo en base a la *performance*, obteniendo la importancia de cada de cada variable. Finalmente, se utilizará el modelo para clasificar a las nuevas inscripciones en corta o larga duración, considerando a los vehículos producidos por las organizaciones agrupadas en la ADEFA durante el año 2021. Agrupando los resultados por modelo de vehículo, se obtendrá la proporción de bajas anticipadas esperadas, de darse la baja. Se realizará una ponderación de estos resultados por la proporción de vehículos que se dan de baja en cada modelo y, de esta forma, se obtendrá la cantidad de vehículos que se espera que tengan una baja anticipada.

1. La gestión de datos en el sector automotor argentino

En este apartado se presentan dos organizaciones responsables de la gestión de datos en el mercado automotor argentino. Luego, se establecen las diversas fuentes de datos a utilizar y los responsables del mantenimiento de cada una de ellas. Se analiza la importancia de una gestión adecuada de los grandes volúmenes de datos en organizaciones para mejorar la toma de decisiones. Considerando aspectos mencionados por la literatura sobre la demanda de automóviles, se evalúa el caso específico de Argentina. Para ello, se presenta un análisis de la evolución de las inscripciones y bajas en registros seccionales.

1.1. El rol de las organizaciones en la tenencia automovilística: el caso de ADEFA y la DNRPA

La Asociación de Fábricas de Automotores (ADEFA) agrupa a doce terminales automotrices que producen automóviles en la Argentina: vehículos utilitarios livianos y pesados; además de componentes de transmisión, cajas y motores. En la Tabla A1 del Apéndice A se presentan las organizaciones asociadas, incluyendo los vehículos que se producen en el año 2021. Se observa que la mayoría de las empresas se concentran en Buenos Aires, Córdoba y Santa Fe. En este sentido, la producción de automotores de origen nacional pareciera no estar dispersa. En este punto es necesario destacar que los automotores registrados en Argentina provienen de tres orígenes distintos: nacionales, importados o vinculados al Protocolo 21[3].

Esta organización agrupa información relevante del mercado automotor, principalmente relacionada a la producción y venta de vehículos. Los datos vinculados a la evolución de ventas y producción de distintos tipos de vehículos no cuentan con un portal de datos abiertos, sino que se disponibilizan mediante informes anuales. Esto implica un problema importante en relación con el acceso a la información. Existen organizaciones que buscan generar una mayor apertura de datos, como es el caso de Alphacast, que cuenta con una API (*Application Programming Interface)* para acceder a una gran cantidad de datos de coyuntura económica. Por esta razón, a pesar de que la fuente de datos es la ADEFA, se hará uso de esta API para obtener los datos correspondientes a la oferta de la cadena automotriz argentina.

[3] Los automotores correspondientes al Protocolo 21 son importados, pero reciben el mismo tratamiento que los automotores nacionales: arancel, normas tributarias, entre otras.

Las organizaciones agrupadas en la ADEFA representan la oferta existente de vehículos nacionales para abastecer la demanda de los ciudadanos argentinos. Al analizar la demanda, una opción es considerar los vehículos inscriptos en los registros seccionales del automotor, distribuidos en todo el territorio argentino. Estos registros recopilan información de cada una de las inscripciones, transferencias o bajas de vehículos. Los datos son enviados a la Dirección Nacional de los Registros Nacionales de la Propiedad del Automotor (DNRPA), entidad encargada de la regulación de la propiedad automotor. Dentro de las facultades de la DNRPA se encuentra la realización de estudios estadísticos sobre el movimiento registral jurisdiccional y centralizado.

Al estar considerando datos recopilados por una organización estatal, es importante mencionar el contexto en el cual se desenvuelve esta entidad. El Diagrama A1 del Apéndice presenta un organigrama del Ministerio de Justicia y Derechos Humanos, organismo en el cual se encuentra incluida la DNRPA. Tal como se observa, la Dirección de Registros Seccionales corresponde a una subdirección de la DNRPA.

Una de las problemáticas asociadas al contexto de grandes volúmenes de datos es la agregación de diversas fuentes. La DNRPA es la entidad encargada de agregar los datos recopilados por cada uno de los registros seccionales. En la Figura A1 del Apéndice A se presenta la distribución provincial de los registros seccionales. Se observa que las provincias de Buenos Aires, Córdoba y Santa Fe concentran una gran cantidad de registros seccionales, lo cual es similar a la ubicación de las empresas pertenecientes a la ADEFA. Sin embargo, en el resto de las provincias la cantidad de registros oscila entre 5 y 35, una cantidad considerable. Esta dispersión geográfica y administrativa de los registros del automotor trae aparejados diversos problemas. En principio, la homologación del uso de nomenclaturas, tal como se mencionará en la siguiente sección, es un punto para tener en cuenta dada la relevancia de los datos.

El hecho de que la gestión de datos en el mercado automotor se encuentre dispersa en diversas organizaciones presenta una oportunidad grande: combinarlas podría ser una herramienta importante para mejorar la toma de decisiones. En particular, entender la demanda desde el punto de vista de los registros automotores para hacer más eficiente la oferta. De esta forma, mediante el uso de datos de la DNRPA se intentará establecer un método adecuado para la gestión de datos que permita asignar más eficientemente los recursos para la producción de autopartes por parte de las organizaciones agrupadas en la ADEFA.

1.2. El impacto de los grandes volúmenes de datos en las organizaciones

El análisis de datos ha experimentado un notable crecimiento en los últimos años, principalmente debido al desarrollo de nuevas tecnologías que permiten almacenar y analizar grandes volúmenes de datos a un bajo costo y de manera accesible. En el contexto actual, los datos se han convertido en un activo indispensable para las organizaciones, con lo cual la gestión de datos constituye una de las principales problemáticas a abordar.

A pesar de que en la literatura no existe unanimidad en la definición de "*Big data",* hay consenso sobre el impacto disruptivo que trae aparejado el uso de grandes volúmenes de datos. Esto es, la necesidad de captura, almacenamiento y análisis por parte de las organizaciones (Joyanes, 2016). El uso de *"Big data" o "*grandes datos" para referirse a este nuevo paradigma no debe confundirse con gran volumen de datos, sino con un nuevo esquema en donde los datos muchas veces no son generados mediante las formas tradicionalmente utilizadas. En este sentido, en *Big data* se incorporan nuevas fuentes, que anteriormente no existían y que pueden ser de gran utilidad para mejorar la eficiencia en la toma de decisiones en las organizaciones.

Siguiendo a Mergel, Rethemeyer e Isett (2016), es posible establecer tres fuentes de datos en "*Big data":* (1) datos creados por individuos en sus interacciones online, (2) datos generados mediante sensores y, por último, (3) datos generados por organizaciones en su operatoria. Dentro del primer grupo se incluyen los datos generados en redes sociales, una fuente que ha tomado gran relevancia en los últimos años. Los datos generados en la operatoria habitual de las organizaciones suelen mantener un formato estructurado y ordenado, aunque esto no significa que sean menos susceptibles de problemas de gestión en relación con el resto de las fuentes.

La aparición del concepto de *Big data* trajo aparejadas numerosas implicancias en la arquitectura de negocios de las organizaciones públicas y privadas. Desde un enfoque de las organizaciones privadas, el uso de *Big data* permite incrementar el valor económico de las organizaciones, tanto en términos monetarios como en participación en el mercado. Por otro lado, la gestión adecuada de grandes volúmenes de datos en el sector público permite tomar decisiones más eficientes en búsqueda del bienestar de la sociedad.

En este contexto, existen una serie de potenciales problemas asociados al uso de *Big data*. En primer lugar, al estar trabajando con grandes volúmenes de datos que muchas veces contienen información sensible, la seguridad de los datos constituye un elemento fundamental. Dado que la privacidad de los individuos incluidos en las bases es primordial, se deben utilizar técnicas de anonimización de cada uno de los registros. Este proceso debe tener en cuenta la generación de identificadores únicos por individuo en el caso en que existan bases

relacionales, en donde un individuo puede aparecer más de una vez. Por otro lado, cumplir con ciertos estándares de nomenclaturas en las variables categóricas permite realizar cruces de datos y, de esta forma, minimizar los errores y el costo del análisis. Sumado a esto, el análisis de este tipo de datos requiere de una gran capacidad de cómputo y aprendizaje continúo de nuevas herramientas para manipular volúmenes de datos que cada vez son mayores.

Al hablar de *Big Data* es necesario hacer referencia al ciclo de vida de los datos. Inicialmente los datos son capturados por una organización, en este caso por cada uno de los registros seccionales en cada trámite. Luego, es la DNRPA la organización que los agrega y almacena. Existe un procesamiento de datos que agiliza la toma de decisiones por parte de esta entidad mediante la utilización de los datos. La publicación de este tipo de datos mediante un portal de datos abiertos permite pensar en ellos como un bien público: existe un acceso equitativo por todos los actores de la sociedad. Al vincular *Big Data* con los algoritmos de aprendizaje automático es fundamental entender que la calidad de los modelos está estrechamente vinculada a los datos de entrenamiento. En este sentido, los datos de entrenamiento deben ser similares a los datos para los cuales se desea generar predicciones ya que de no ser así el modelo perdería capacidad predictiva. Polyzotis et al. (2018) presentan un resúmen de los principales desafíos en en cada una de las etapas del ciclo de vida de los datos.

En este trabajo se hará uso de un gran volumen de datos generados por una organización (pública) para mejorar la eficiencia en la toma de decisiones de otra organización (privada). Estos datos se generan con la finalidad de mantener un registro de la operatoria a nivel nacional, pero constituyen una fuente de datos útil para otras actividades.

1.3. Situación de la demanda de vehículos para uso personal

La tenencia de automóviles tiene un rol fundamental en las decisiones de desplazamiento diario individual y de los hogares, influenciando decisiones de corto y largo plazo (Anowar et al., 2014). Con el crecimiento de las ciudades, el estudio de la movilidad urbana ha adquirido gran relevancia en los últimos años, entendiendo a la movilidad como un concepto más amplio que el de transporte. Gentile y Nökel (2016) mencionan que “el objetivo de la movilidad es el acceso, no el movimiento”, dado que lo que se busca es permitir el acceso

a un lugar o servicio mediante el uso de un medio de transporte. De esta forma, no contar con un medio de transporte adecuado es un impedimento para acceder a derechos humanos básicos[4].

En el diseño de políticas públicas vinculadas al transporte se analizan dos frentes: el transporte público y el transporte privado. Diversos estudios analizan el rol de los acuerdos público-privados para la provisión de infraestructura que permita la correcta circulación de vehículos (Tsamboulas et al., 2013). La cadena automotriz-autopartista es uno de los principales sectores de la industria argentina, con una relevancia estratégica en términos de producción, empleo y exportaciones. La importancia de este sector en la economía hace que las políticas de incentivo sean necesarias. Canatrella et al. (2008) sugieren que "El fortalecimiento de la industria autopartista se hace necesario ante la magnitud del déficit de la industria automotriz" (p. 40).

En la Figura B1 del Apéndice B se presenta la evolución en las ventas y producción de automóviles. Los datos de producción representan la dinámica de la oferta de vehículos nacionales. Se observa una dinámica similar en ambas series, con una diferencia considerable entre los años 2016 y 2018, siendo la cantidad de ventas superior a la producción. Luego de mediados de 2018, las ventas comienzan a decrecer, llegando a un mínimo en el comienzo de la restricción obligatoria ocasionado por la pandemia. Con la apertura de actividades, comienza a recuperarse el mercado.

Para hacer frente a las necesidades de la población en términos de transporte privado, es necesario entender los factores que afectan la demanda de automóviles. La tenencia de automóviles, y las dimensiones asociadas con el tipo de vehículos y su uso, es uno de los tópicos de transporte más estudiados en términos de demanda (Anowar et al., 2014). En estudios anteriores, el foco solía estar en el análisis de la tenencia a nivel agregado. La Figura B2 del Apéndice B presenta las inscripciones y bajas realizadas, según el origen de los vehículos. Estos datos provienen de la DNRPA e identifican la demanda de vehículos dentro de Argentina.

El enfoque anterior no captura el mecanismo que realmente guía el proceso de decisión individual. En el contexto actual de *Big Data*, es posible realizar análisis más detallados dada la gran cantidad de datos disponibles. Se considera como primer punto de investigación a los factores individuales, vinculados a las características de cada persona. Este es el caso típico de la teoría de la decisión, una disciplina que analiza cómo elige una persona aquella acción que, de entre un conjunto de acciones posibles, le conduce al mejor resultado dadas sus preferencias (González, 2004).

[4] Contemplados en La Declaración Universal de Derechos Humanos.

Aspectos como la estructura familiar, ubicación, reconocimiento de la marca o confiabilidad en el vehículo influyen sobre las preferencias de cada individuo[5] (Lee y Govindan, 2014). Incorporar cuestiones geográficas en el análisis es necesario por dos razones: la calidad del transporte público y las distancias a recorrer en la vida diaria. La competitividad de los automóviles en relación con otros medios de transporte es más elevada en los desplazamientos en zonas periféricas debido a la menor densidad de demanda (Müller, 2013; Clark, 2016). Tal como se observa en la Figura B3 del Apéndice B, provincias con mayor densidad poblacional tienen más inscripciones y bajas, lo cual es esperable. Estas provincias también son las provincias en donde se encuentran más registros seccionales.

Existen otro tipo de modelos de decisión sobre la tenencia de vehículos que analizan el ciclo de vida de las personas. Estos modelos son dinámicos e incorporan el impacto de eventos como el nacimiento de un hijo, mudanza o cambio de empleo sobre la adquisición, duración o reemplazo de la tenencia de un automóvil (Yamamoto, 2008). A pesar de ser un enfoque más sofisticado en cuanto al análisis de las decisiones, es más complejo de analizar.

Por otro lado, el contexto también tiene influencia sobre las preferencias individuales. El marco regulatorio y la situación económica deben considerarse para entender las decisiones desde una perspectiva temporal. Dentro de los aspectos económicos se incluye el costo del combustible, variable fundamental en la decisión sobre el uso de automóvil propio. Numerosos estudios analizan el efecto de las regulaciones sobre las decisiones de transporte de las personas (Sayyadi y Awasthi, 2017; Guo, 2013). Otro aspecto del contexto es la situación socioeconómica de los vecinos, donde la tenencia está influida por las decisiones del grupo al que se pertenece (Goetzke y Weinberger, 2012).

Un último punto para mencionar con relación al transporte es la conciencia ambiental. “El sector representa el 23 % de las emisiones mundiales de gases de efecto invernadero relacionadas con la energía” (Banco Mundial, 2021). En los últimos años se han impulsado ideas sobre el cuidado del medioambiente, incentivando el uso del transporte público o medios alternativos como las bicicletas. Esto reduce el uso de los vehículos propios, pero no necesariamente la tenencia. Se estima que gran parte de la valoración de tener un vehículo propio no proviene de su uso, sino del estatus social asociado a ser propietario, junto con la opción de poder viajar a donde se quiera en el momento en que se quiera (Moody et al., 2021).

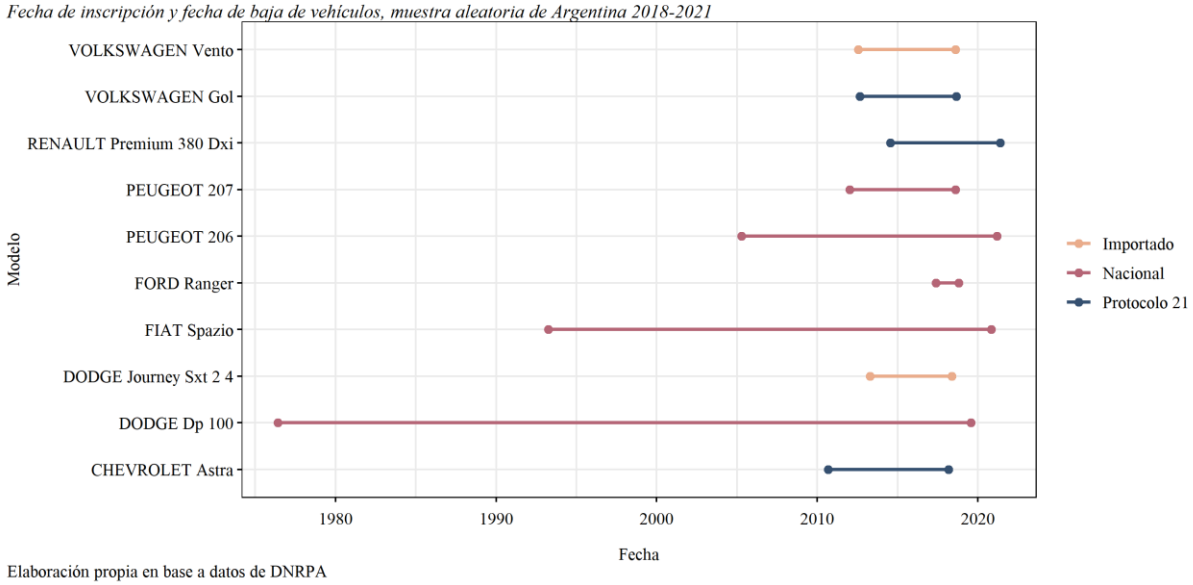
El entendimiento de los patrones que determinan la demanda de automóviles es fundamental para tomar medidas de política de transporte adecuadas. El sector automotor en

[5] Considerar a la ubicación como una variable exógena puede no ser adecuado, dado que podría argumentarse que las preferencias de cada persona son las que determinan su zona de residencia.

Argentina ha sido analizado por diversos autores desde un enfoque de la oferta (industria), sin embargo, no existen muchos estudios sobre la demanda. En la Figura 1 se presenta una muestra aleatoria de 10 observaciones de vehículos que han sido dados de baja entre los años 2018 y 2021. Tal como se observa, ciertos vehículos han tenido una duración considerablemente superior en años. El objetivo de este trabajo es entender los factores que explican estas diferencias en tiempo de tenencia de los vehículos para uso personal.

Figura 1

Fecha de inscripción y fecha de baja de vehículos, muestra aleatoria de Argentina 2018-2021



Elaboración propia en base a datos de DNRPA

Un aspecto importante para tener en cuenta al analizar el tiempo que transcurre entre la inscripción y la baja de un vehículo es el origen: nacional o importado. La Figura B4 del Apéndice B indica que existe evidencia para considerar que existen diferencias significativas en la duración según el origen de los vehículos. Esta es una cuestión relevante, dado que el foco del trabajo es el fortalecimiento de la industria automotriz nacional mediante el análisis de datos del mercado automotor.

El entendimiento del contexto en el cual se desenvuelve el mercado automotor argentino, junto con diversos aspectos mencionados por la literatura de demanda automotor, sirven como punto de partida para responder el interrogante planteado. Al estar trabajando con grandes volúmenes de datos, de diversas fuentes, es fundamental hacer un seguimiento detallado de la gestión de los datos a analizar. En el siguiente apartado se hará una explicación metodológica para luego proceder a estimar el efecto de diversas variables sobre la duración de la tenencia.

2. Métodos predictivos aplicados a la duración de la tenencia de automóviles

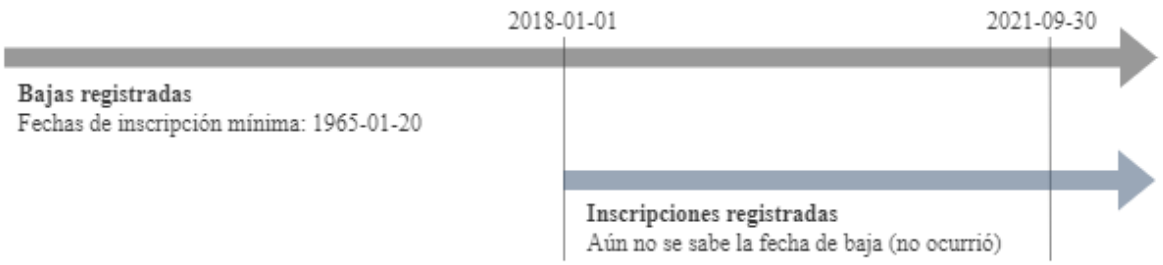
Con el objetivo de determinar el efecto de distintos factores sobre la duración de la tenencia de vehículos propios en Argentina, se propone una metodología de análisis. Dada la distribución de la duración, se decide utilizar un modelo de clasificación para estimar la probabilidad de una baja anticipada. Se presentan las definiciones metodológicas, modelos a utilizar y métricas a evaluar. Se realizan las transformaciones necesarias para la posterior aplicación de modelos estadísticos multivariados o de aprendizaje automático. Se define un modelo *baseline* y luego se ajustan los modelos restantes. Se realiza una búsqueda de hiperparámetros bayesiana en relación con una métrica seleccionada. Finalmente, se selecciona el mejor modelo realizando una evaluación en la partición de evaluación.

2.1. Metodología para clasificar a los vehículos en corta y larga duración

Los datos provistos por la DNRPA incluyen 142.307 bajas y 1.919.091 inscripciones registradas en el período 2018-2021. En el Diagrama 1 se presenta la estructura de los datos. A pesar de que los datos de bajas corresponden al período entre los años 2018 y 2021, la fecha de inscripción inicial del vehículo en general es anterior.

Diagrama 1

Estructura de los datos disponibilizados por la DNRPA



Fuente: Elaboración propia en base a datos de la DNRPA

En la Tabla 1 se presentaron los vehículos relevantes para la ADEFA: automotores que se producen en el año 2021 por las organizaciones de la industria automotriz nacional. Para el caso de análisis, se considerarán estos vehículos, incorporando un segmento adicional que cumple con ciertas condiciones. Se han definido como vehículos relevantes a los modelos con más de 2000 inscripciones entre 2018 y 2021. Además, se consideró la mínima fecha de inscripción de los vehículos con baja: posterior al año 2000 y anterior al año 2013. Esto permite

considerar vehículos que han sido relevantes en el pasado y lo continúan siendo al momento de realizar este trabajo.

Tabla 1

Vehículos relevantes: fabricación nacional actual o relevancia en términos de cantidad de vehículos

Fabricación Nacional Actual	Modelo de vehículo
Sí	308, 408, AMAROK, BERLINGO, C4, CRUZE, FRONTIER, HILUX, HRV, KANGOO, LOGAN, PARTNER, RANGER, SANDERO, SPRINTER
No	208, DUSTER, ECOSPORT, ETIOS, FELINE, FIESTA, FIORINO, FIT, FOCUS, FOX, GOL, KA, MARCH, ONIX, PALIO, POLO, PRISMA, SAVEIRO, SPIN, STRADA, SURAN, TIGGO, TRACKER, VENTO, VOYAGE

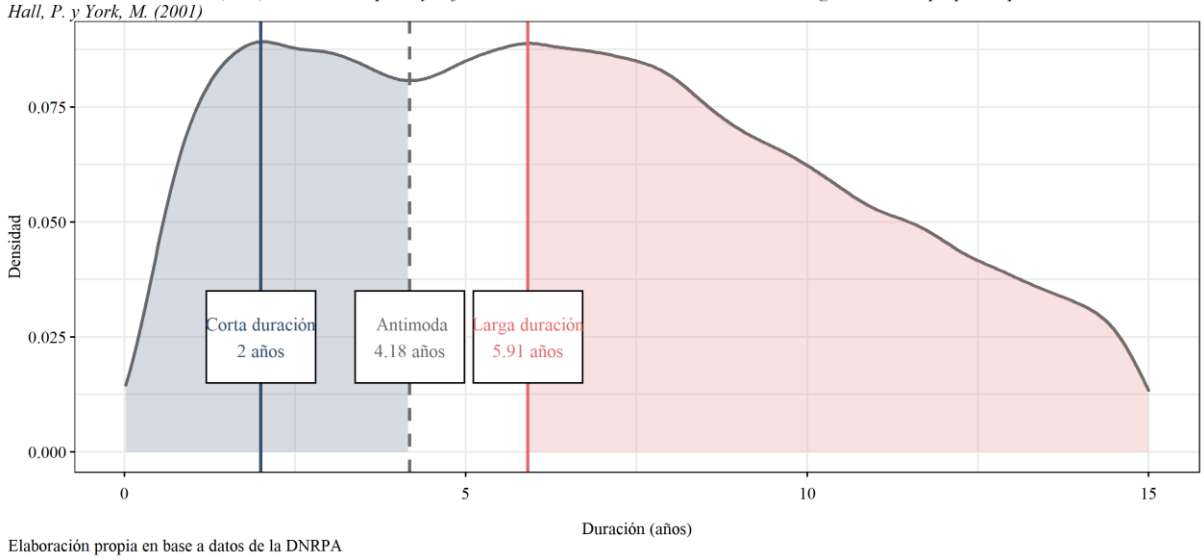
Fuente: Elaboración propia en base a datos de DNRPA, 2021

Inicialmente se buscaba generar un modelo de regresión para estimar la duración de un vehículo, pero los resultados no fueron concluyentes. En este sentido, se observa que la duración de la tenencia presenta una distribución bimodal. Para evaluar la existencia de más de una moda, se utilizó el método propuesto por (Hall & York, 2001). Dado un nivel de significatividad del 5% fue posible rechazar la hipótesis nula de una única moda en favor de la alternativa de existencia de múltiples modas.

En la Figura 2 se presenta la distribución de la duración de la tenencia. Considerando esta distribución, se busca evaluar la existencia de factores que incrementan la probabilidad de que un vehículo tenga una baja temprana, es decir, duración menor que un umbral. Para mejorar la clasificación del modelo, se decide utilizar 2 umbrales: si la duración es menor a cierto umbral se considera que el vehículo tuvo una baja temprana, si la duración es superior a cierto umbral se lo considera de larga duración. Se excluyen del análisis los vehículos con una duración intermedia.

Figura 2

Distribución de la duración (años) entre la inscripción y baja de un vehículo. N=61293. Modas obtenidas según el método propuesto por Hall, P. y York, M. (2001)



Elaboración propia en base a datos de la DNRPA

En total se cuenta con 45.524 observaciones: 35% vehículos de corta duración y los restantes de larga duración. Mediante la implementación de modelos de aprendizaje automático, se busca generar un modelo que permita estimar si los nuevos vehículos inscritos tendrán una baja anticipada. En este sentido, se hará énfasis en los vehículos que al año 2021 están siendo producidos por las empresas incluidas en la ADEFA, presentados en la Tabla A1. En la Tabla C1 del Apéndice C se presenta la descripción de las variables junto con su rol en el modelo. Se consideran tres agrupamientos de los predictores: características del titular, del automóvil y del contexto.

$$P(Y = corta\ duración) = f(X_{automotor}, X_{titular}, X_{contexto}) \quad (1)$$

En relación con aspectos del automotor, se cuenta con información sobre el modelo, marca, categoría, tipo y uso. Además, en la siguiente sección se hará uso de técnicas estadísticas para caracterizar a los vehículos y así obtener variables adicionales: componentes principales y segmentos. Los aspectos vinculados al titular del vehículo se refieren a variables del titular inicial, independientemente de si la titularidad del vehículo ha sido transferida. Esto presenta la ventaja de que es una variable comparable en los registros de bajas e inscripciones y, por lo tanto, se puede utilizar para estimar la probabilidad de baja temprana de una nueva inscripción. Sin embargo, el hecho de que cambie la titularidad podría ser un factor influyente sobre la baja del vehículo y en este caso corresponde a una variable no observable dado que no se cuenta con un identificador único de los vehículos que permita cruzar los datos con la información correspondiente a transferencias de titularidad. Finalmente, se han incluido variables vinculadas al contexto en el cual se ha registrado el vehículo. De esta forma, se incorporan aspectos relacionados con la zona: población, robos, porcentaje de bajas, entre otros.

Para obtener estas variables ha sido necesario realizar un exhaustivo proceso de limpieza de datos, dado que existía una gran cantidad de inconsistencias en los datos. En primer lugar, se contaba con datos de la ubicación del titular sin ningún tipo de estandarización en las nomenclaturas utilizadas. Para solucionar este problema, se decidió utilizar la ubicación del registro seccional en el cual se realizó el trámite. En este sentido, se asume que el trámite fue realizado en las cercanías de la zona de residencia del titular. Por otro lado, las variables vinculadas a los vehículos tampoco contaban con un formato estándar, con lo cual fue necesario realizar un proceso de estandarización de modelos y marcas. En términos de la edad del titular, se encontraron registros inconsistentes, como personas físicas de muy baja edad, los cuales se

decidió imputar con la mediana. Los aspectos vinculados al contexto se incorporaron desde diversas fuentes: Instituto Nacional de Estadísticas y Censos (INDEC), Centro de Estudios Distributivos Laborales y Sociales (CEDLAS) y otras bases de la DNRPA. En este sentido, se utilizaron datos demográficos provenientes del INDEC, junto con datos de pobreza estimados por el CEDLAS. Para incorporar aspectos sobre robos de vehículos se utilizó la base de robos de vehículos provista por la DNRPA.

Debido a que las inscripciones se han realizado en distintos momentos, se asume que no existen efectos que cambian en el tiempo. Es decir, el efecto de una variable sobre la probabilidad de baja temprana es el mismo para un vehículo inscripto en un año como lo es para un automóvil inscripto en otro momento del tiempo. Este es un supuesto muy fuerte, pero ha sido necesario para realizar el análisis, dado que solo se posee registros de bajas realizadas entre 2018 y 2021. En relación con este aspecto, es necesario destacar que los en los vehículos considerados para el análisis el año mínimo de inscripción considerado es el 2003.

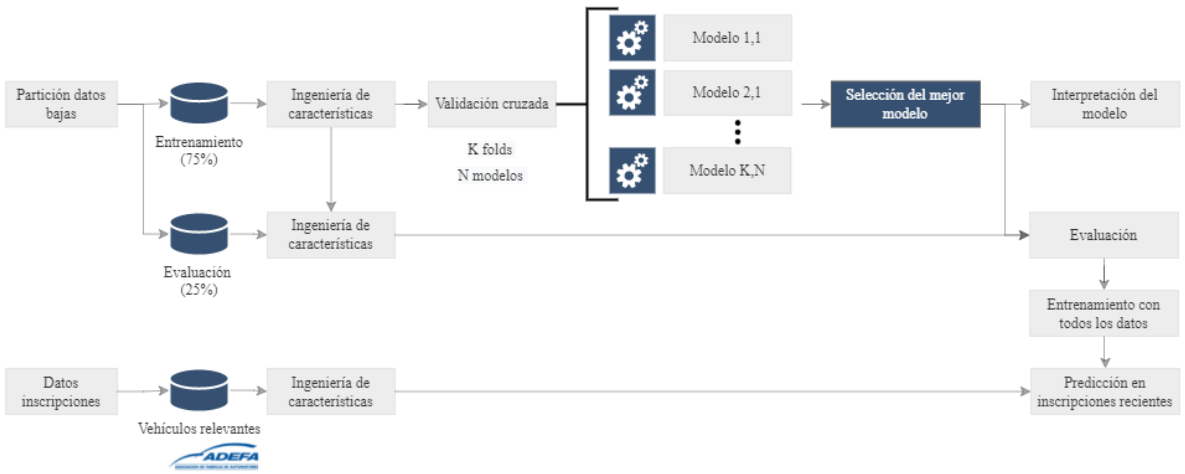
Se toma como modelo base (*baseline)* un modelo de regresión logística. Luego, se procede a ajustar modelos de aprendizaje automático. Para el ajuste de modelos, se considera el método de *"hold-out cross validation"*. En este método, se define una partición inicial de los datos en entrenamiento y evaluación. Se busca obtener los mejores hiperparámetros mediante el método de validación cruzada y luego evaluar la capacidad predictiva sobre una partición de evaluación que no haya sido utilizada en el ajuste de los modelos. En este sentido, suele ser una técnica muy utilizada para evitar el sobreajuste.

En el Diagrama 2 se presenta un esquema de la metodología a utilizar. En primer lugar, se realiza un preprocesamiento de los datos. Este proceso incluye la limpieza de los datos de texto (modelo, marca, categoría). Luego, se añaden variables adicionales vinculadas al contexto (principalmente relacionadas con características de la ubicación en donde ocurrió el registro). Se genera una partición de los datos de bajas en dos conjuntos de datos: partición de entrenamiento (70% de las observaciones) y partición de evaluación (30% de las observaciones). Se utiliza la partición de entrenamiento para generar la ingeniería de variables necesaria para la aplicación de diversos modelos de aprendizaje automático. En general, este tipo de modelos requieren variables numéricas y normalizadas. Para evitar utilizar los datos de evaluación, se utilizará la partición de entrenamiento para generar las transformaciones que luego son aplicadas a la partición de evaluación. Por ejemplo, los promedios y desvíos de las variables, agrupamientos de ciertas categorías en categoría "otro", entre otras, son obtenidas del conjunto de datos de entrenamiento. Luego de contar con los datos en un formato adecuado, se realiza el ajuste de hiperparámetros de diversos modelos con *k-fold cross validation.* Se

visualizan los resultados, obteniendo el mejor modelo según una métrica determinada y su desvío en cada pliegue. El mejor modelo es evaluado en la partición de evaluación. Si su *performance* es aceptable, se realiza el entrenamiento con todos los datos (entrenamiento y evaluación) para luego interpretar el modelo en búsqueda de cuáles han sido los principales factores que permiten clasificar vehículos en baja anticipada o larga duración. Finalmente, obtienen las predicciones sobre el conjunto de datos de inscripciones recientes, considerando los vehículos relevantes para la ADEFA.

Diagrama 2

Resúmen metodológico para el ajuste de modelos sobre la clasificación de vehículos en corta y larga duración.



Fuente: Elaboración propia

Los modelos utilizados son: *elastic net* (Friedman, Hastie y Tibshirani, 2010), árbol de decisión (Breiman, Friedman, Olshen y Stone , 2017), *random forest* (Breiman, 2001), *boosted trees* (Chen y Guestrin, 2016), *bagged trees* (Breiman L., 1996), *support vector machines* (Chang y Lin, 2011). Para el ajuste de modelos se utilizará el *framework* de *Tidymodels*, (Kuhn y Wickham, 2020). Dado que es un problema de clasificación, se evaluarán las siguientes métricas: *accuracy*, *recall*, área bajo la curva ROC, precisión, *f_meas*. La *accuracy* es relevante ya que predecir bien ambas clases es relevante. Sin embargo, dado que es un problema desbalanceado, donde la clase de corta duración cuenta con un porcentaje menore de observaciones, la *recall* es una métrica apropiada para evaluar si el modelo logra clasificar ese tipo de observaciones. Para tener en cuenta la precisión y la *recall* simultáneamente, se ha incorporado al f_meas como métrica adicional. Esta métrica considera la media armónica entre la *recall* y la precisión. A continuación, se muestra la ecuación:

$$F_\beta = (1+\beta^2)\frac{precision * recall}{(\beta_2 * precision) + recall} \qquad (2)$$

En donde, si se considera $\beta = 1$ se obtiene el conocido *F1-score*. En este caso, es esta la métrica que decidió evaluarse, es decir, se ha considerado $\beta = 1$. Dado que se cuenta con menos observaciones de corta duración que de larga duración, se han evaluado dos alternativas: utilizar técnicas de sobre o submuestreo y optimizar la métrica de *recall* mediante el ajuste de hiperparámetros. Ambas técnicas han sido exitosas, aunque se decidió utilizar la técnica de optimizar la *recall* para no distorsionar la distribución de las observaciones según clasificación.

2.2. Análisis exploratorio y procesamiento de datos

Previo al ajuste de modelos de clasificación, se realiza un análisis exploratorio de los datos para entender los aspectos que podrían mejorar la capacidad predictiva del modelo. El análisis está organizado en las tres categorías de variables descriptas en la Tabla C1. En primer lugar, se aplicarán métodos de reducción de dimensionalidad y segmentación de los automotores. Luego, se analizará la distribución de las variables de los titulares o el contexto en relación con la clasificación.

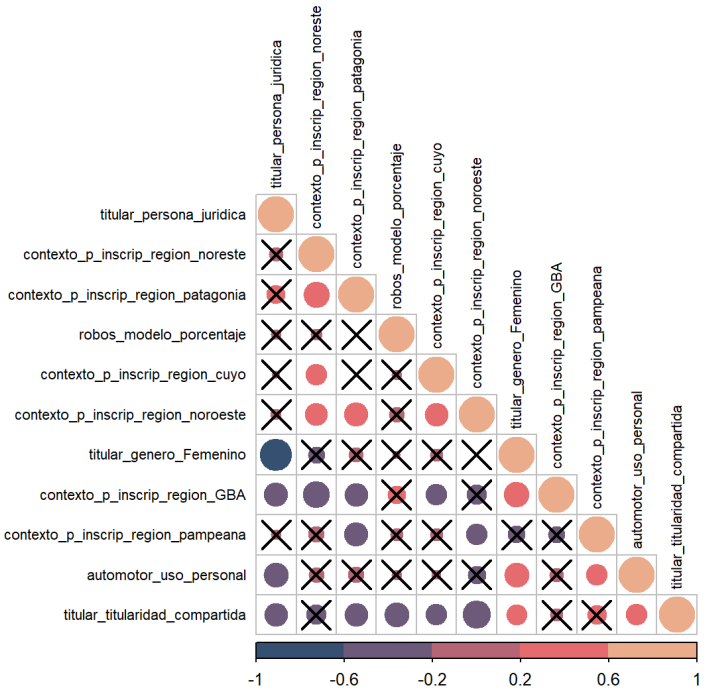
2.2.1. Automotores

En el caso de los automotores, se aplica la técnica de análisis de componentes principales (PCA) con el objetivo de obtener una representación numérica que identifiqua a cada modelo de vehículo, dado que incorporar los N vehículos como variables *dummy* puede ser problemático. Para el análisis de componentes principales se utilizan variables vinculadas a cada modelo de automóvil, al titular (porcentaje de género femenino, personas jurídicas, titularidad compartida o extranjeros), ubicación (porcentaje de inscripciones en cada región), robos (por modelo y por marca), entre otras. En la Figura C1 del Apéndice C se presenta la distribución de ciertas variables asociadas a cada uno de los modelos de vehículos seleccionados.

Se genera la matriz de correlación (R), presentada en la Figura 3. Las correlaciones que contienen una cruz no son estadísticamente significativas considerando un nivel de significatividad del 10%. El Test de Bartlett verifica si la matriz de correlación se diferencia significativamente de la matriz identidad. Realizar Análisis de Componentes Principales no tendría sentido ante la inexistencia de correlación entre las variables. En este caso, es posible rechazar la hipótesis nula en favor de la hipótesis alternativa de existencia de correlación entre las variables y, por esta razón, se procede a realizar el análisis de componentes principales.

Figura 3

Matriz de correlación de las variables seleccionadas para análisis de componentes principales de los vehículos



Fuente: Elaboración propia

El análisis de componentes principales tiene como objetivo representar la información de N observaciones y P variables en un número menor de variables construidas como una combinación lineal de las variables originales. Este método permite la interpretación de un gran número de variables a través de componentes no correlacionados entre sí. Matemáticamente, el primer componente se escribe como la combinación lineal de las variables que captura la mayor variabilidad posible de los datos. Las siguientes componentes deben explicar la mayor variabilidad posible restante que no haya sido explicada por las componentes anteriores. Es decir:

$$Z_1 = \varphi_{11}X_1 + \varphi_{21}X_2 + \cdots + \varphi_{p1}X_p \qquad (3)$$

Siendo Z_1 el primer componente principal, construido como una combinación lineal de las variables $\{X_1, X_2, \ldots, X_p\}$, donde $\{\varphi_{11}, \varphi_{21}, \ldots, \varphi_{p1}\}$ son las cargas factoriales (*loadings*) de

dicho componente. Las cargas factoriales pueden interpretarse como el peso que tiene cada variable en cada componente y, por lo tanto, permiten entender qué tipo de información recoge cada uno. Previo a la aplicación del método, se centran y escalan los datos para evitar que se tomen más relevancia a las variables con valores mayores (mayor varianza). La Tabla C2 del Apéndice C presenta el porcentaje de variabilidad explicada por cada componente. Tal como se observa, las primeras dos componentes logran explicar más de un 50% de la variabilidad de los datos. Mediante el método de Horn's Parallel Analysis se seleccionan 4 componentes, según los autovalores ajustados mayores a 1. En este sentido, incorporar estos componentes al modelo permitirá incorporar información vinculada a cada uno de los vehículos contemplados en los datos, en un formato numérico. Esto permite evitar el uso de las 39 variables *dummy* que deberían utilizarse por modelo de vehículo, las cuales hubieran sido binarias y no contemplarían la similitud existente entre distintos modelos.

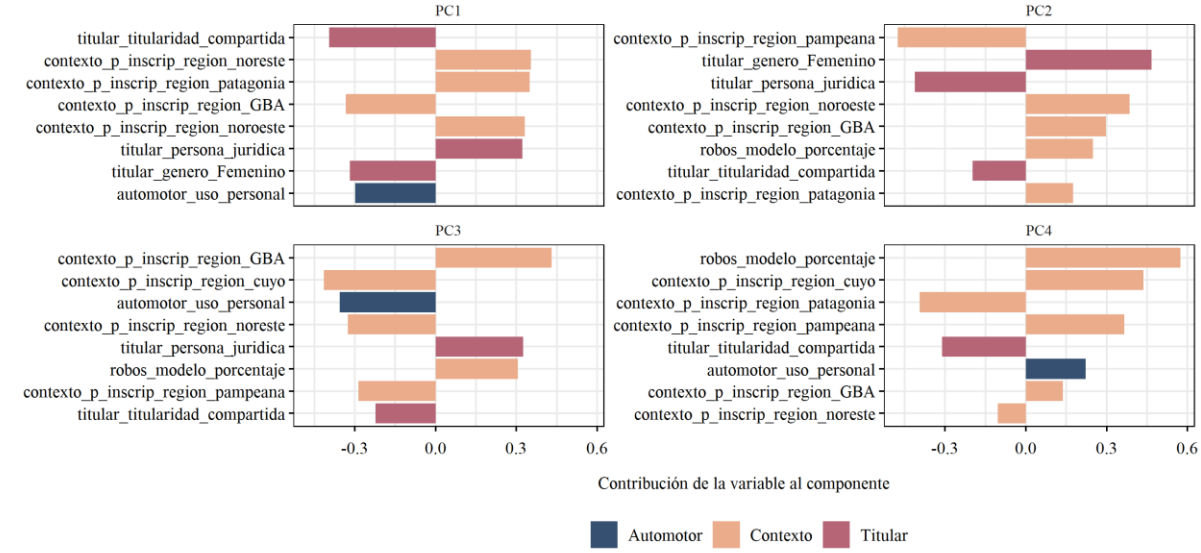
La correlación de cada variable con un componente principal permite entender qué representa cada componente. La carga factorial de la variable j en el componente h, L_{jh} está dada por:

$$L_{jh} = \frac{Cov(X_j, Z_h)}{Var(X_j)^{\frac{1}{2}} Var(Z_h)^{\frac{1}{2}}} \qquad (4)$$

En la Figura 4 se visualizan los primeros 4 componentes, donde el signo de las cargas de cada variable en cada componente mantiene lo observado en la matriz de correlación. El primer componente indica que la titularidad compartida es una variable que explica gran parte de la variabilidad entre vehículos. Los vehículos con alto porcentaje de inscripciones en las zonas Noreste y Patagonia zonas no tienen un elevado porcentaje de titulares que compartan la titularidad del vehículo. Por otro lado, el componente 2 se ve afectado negativamente por el porcentaje de vehículos de determinado tipo en la región Pampeana y positivamente por el porcentaje de mujeres que han realizado inscripciones de ese modelo. ´

Figura 4

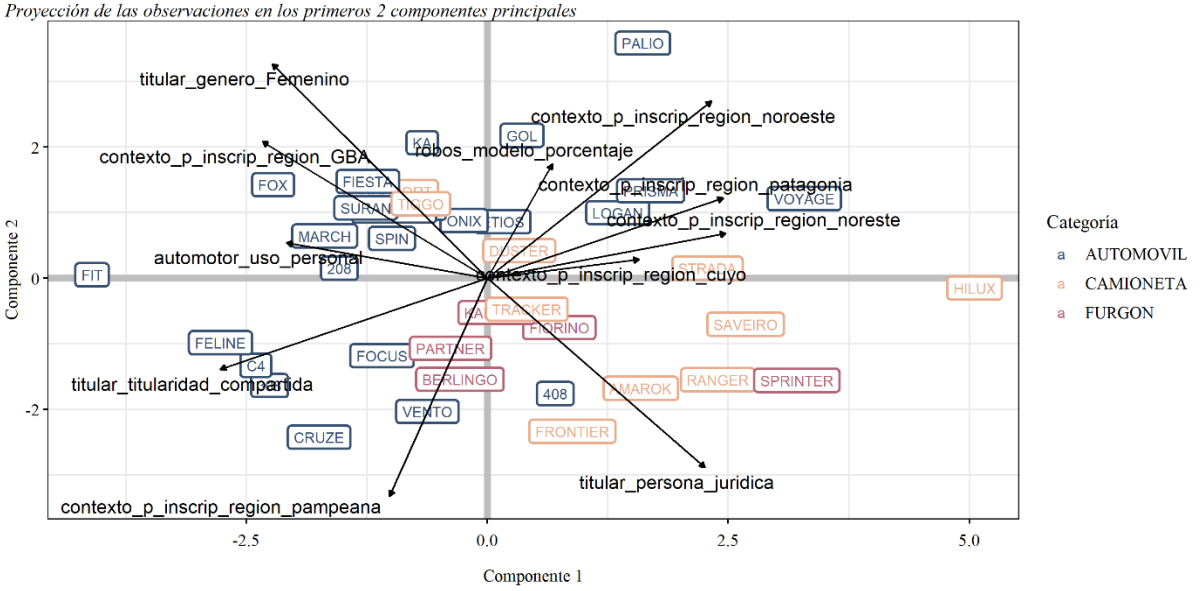
PCA: Variables vinculadas a automotor



Proyectando los vehículos en el plano de los dos componentes seleccionados es posible observar qué vehículos son más similares entre sí. La Figura 5 muestra los resultados, donde los colores representan la categoría de vehículo, variable que no ha sido incluida en el análisis de componentes principales. Se observa que, a pesar de no haber incluido la variable de categoría de vehículo en el análisis de componentes, las categorías se agrupan cerca, principalmente la categoría “fugron”. La decisión de no incluir la categoría en este análisis se debe a que sólo existen 3 categorías y, por ende, es una variable que puede incluirse en los modelos de aprendizaje automático sin necesidad de agruparla o eliminar niveles.

Figura 5

Proyección de las observaciones en los primeros 2 componentes principales



Además de las variables de componentes principales, se decidió segmentar a los vehículos en N grupos homogéneos entre sí y diferentes del resto. Este enfoque, distinto al anterior, permitirá incorporar en el modelo variables vinculadas a los vehículos agrupados en menos categorías. El método de Ward, o método de varianza mínima, busca minimizar en cada paso el incremento en la suma de errores cuadrados intra-clusters, E, definida como:

$$E = \sum_{k=1}^{K} E_k = \sum_{n=1}^{N} \sum_{p=1}^{P} (x_{kn,p} - \bar{x}_{k,p})^2 \qquad (5)$$

Donde k representa al conglomerado, n a la observación y p a la variable. Se parte de todas las posibles combinaciones y se fusiona el par con mínima distancia. La Figura C2 del Apéndice C representa los agrupamientos generados, donde se ha seleccionado 3 como el número de segmentos óptimo considerando un conjunto de métricas. En la Tabla 2 se presentan los modelos de vehículos incluidos en cada uno de los segmentos. Se observa que los 3 agrupamientos incluyen vehículos de distintas categorías. En este sentido, incorporar la variable de categoría junto con la de segmento es relevante, dado que ambas clasificaciones presentan distinta información.

Tabla 2
Variables

Cluster	Modelo de vehículo
1	208, 308, BERLINGO, C4, CRUZE, FELINE, FIORINO, FOCUS, KANGOO, PARTNER, VENTO
2	408, AMAROK, FRONTIER, HILUX, RANGER, SAVEIRO, SPRINTER, STRADA, VOYAGE
3	DUSTER, ECOSPORT, ETIOS, FIESTA, FIT, FOX, GOL, KA, LOGAN, MARCH, ONIX, PALIO, POLO, PRISMA, SANDERO, SPIN, SURAN, TIGGO, TRACKER

Fuente: Elaboración propia

2.2.2. Titular

En la Tabla C3 del Apéndice C se presentan las diferencias entre clasificación según variables vinculadas al titular del vehículo. Estas variables están asociadas al titular inicial, es decir, si existiera una transferencia de dominio igualmente se considera como variable aspectos asociados al titular inicial. En todos los casos existe evidencia de diferencias significativas entre grupos considerando los *tests* realizados según tipo de variable. En relación con la nacionalidad del titular, en titulares extranjeros es mayor el porcentaje de larga duración: 4.79% titulares

extranjeros con larga duración contra 4.16% extranjeros con corta duración. La titularidad compartida también parece influir positivamente sobre la duración: 6.52% de los titulares de vehículos de larga duración tienen titularidad compartida mientras 5.1% tienen corta duración. El género presenta una distribución similar entre ambos grupos. En este caso, la cantidad de valores faltantes es considerable.

Un aspecto relevante para analizar es la edad del titular al momento de la inscripción. Esta variable está influida por el tipo de titular: si titular de la inscripción inicial es una persona jurídica se considera el año de nacimiento de la entidad, mientras que en personas físicas se considera el año de nacimiento de la persona. En la Tabla C4 del Apéndice C se presenta la distribución según el tipo de persona. Se observa que en el grupo de personas jurídicas existe un mayor porcentaje de vehículos de corta duración. Además, la "edad mediana" es considerablemente inferior (18 años contra 37 años en el caso de personas físicas), con un rango intercuartílico más amplio.

2.2.3. Contexto

Analizando aspectos vinculados al contexto, en la Figura C3 del Apéndice C se presenta el porcentaje de bajas anticipadas sobre el total de bajas por provincia. De esta forma, es posible ponderar la cantidad de bajas anticipadas según la cantidad de bajas que se registran en cada zona. Se observa que en las provincias del norte este porcentaje es considerablemente superior. Esto indica que existen diferencias regionales relevantes para clasificar observaciones en corta o larga duración. Considerando este aspecto, se generó una variable regional para agrupar a las provincias. En la Tabla C5 del Apéndice C se presenta el porcentaje de vehículos de corta duración sobre el total de bajas por región.

En la Tabla C6 del Apéndice C se muestra la distribución de otras variables del contexto: población del departamento en el cual se registró el vehículo, cantidad de robos sobre la población del departamento, porcentaje de personas en situación de pobreza crónica y porcentaje de robos del modelo de automotor. Para la variable de pobreza se utilizaron datos provistos por el CEDLAS, los aspectos de población fueron obtenidos del último Censo (2010) y los robos se obtuvieron de la DNRPA. A pesar de que en algunas variables la interpretación pareciera ser contraintuitiva, como es el caso de la cantidad de robos de vehículos sobre la población de esa zona, se busca entender si la interacción con otras variables es relevante.

2.3. Determinación de los factores que explican la duración de la tenencia de vehículos en Argentina

En esta sección se hará uso de modelos estadísticos multivariados y de aprendizaje automático para establecer un modelo para clasificar las observaciones en corta y larga duración. Se utiliza el método de *hold out cross validation* para lo cual el primer paso necesario es particionar al conjunto de datos en un grupo de observaciones de entrenamiento y otro de evaluación. Para ello, se realiza una partición estratificada según la clasificación, considerando el 70% de las observaciones para entrenamiento y el 30% restante para evaluación de la *performance* luego de haber realizado el ajuste de hiperparámetros.

Previo al ajuste de modelos, se establece un método para generar las variables necesarias para estimar el modelo. Este método incluirá los siguientes aspectos: imputación de valores faltantes mediante la mediana o moda dependiendo del tipo de variable, normalización de variables, generación de variables *dummy* a partir de las variables categóricas y remover variables con baja varianza o altamente correlacionadas (se consideró un umbral de 0.8). Todos estos aspectos se realizan considerando la partición de entrenamiento, es decir, aspectos como la mediana o moda para imputación de valores faltantes provienen de los datos de entrenamiento. Esto permite evitar un problema común en modelos de aprendizaje automático: *data leakage*. En este sentido, se debe evitar que la información fuera de los datos de entrenamiento sea utilizada para construir el modelo. Por esta razón, los datos de la partición de evaluación solo se utilizan en el momento final para evaluar el modelo, no pueden ser utilizados antes.

El uso de modelos de aprendizaje automático permite encontrar patrones ocultos en los datos, como relaciones no lineales, que permiten predicciones considerablemente mejores que modelos más simples. Sin embargo, su uso trae aparejados ciertos problemas. El principal problema es que cuanto más complejo es el modelo es más difícil determinar qué factores influyen sobre la predicción y de qué forma lo hacen. En este sentido, suelen denominarse modelos de "caja negra" o *"black box"*. Por esta razón, el uso de estos modelos debe ser justificado por una considerable mejora en *performance* en relación con modelos más simples y de fácil interpretación. Cuando no es posible explicar las razones por las cuales un modelo predice de la forma que lo hace, es más probable que los tomadores de decisiones decidan no implementarlo. Por esta razón, si un modelo de aprendizaje automático resulta considerablemente superador a un modelo más simple es necesario utilizar técnicas específicas para la interpretación de modelos.

Se decidió estimar un modelo de regresión logística para la clasificación y, luego de obtener las métricas asociadas, este será el modelo *baseline* que se busca mejorar. Si no se pudiera mejorar considerablemente este modelo no tendría sentido utilizar modelos complejos. En la Tabla C7 del Apéndice C se presenta el resultado de 3 modelos de regresión logística ajustados: el primero considera sólo las variables vinculadas al automotor, el segundo incorpora aspectos vinculados al titular y el tercero incluye todas las variables posibles. Se realizó un análisis de la calidad del ajuste mediante el método del pseudo R^2 de McFadden, indicando que el tercer modelo es el más adecuado. En el Apéndice C también se incluye la ecuación del modelo seleccionado para clarificar que se está estimando la probabilidad de corta duración. Tal como se observa en la ecuación, los coeficientes impactan sobre el *odd ratio*. En este sentido, para lograr interpretar el efecto de una unidad adicional de cada variable es necesario realizar la transformación logística.

En la Tabla 3 se presentan las métricas asociadas al mejor modelo de regresión logística. A pesar de que la *accuracy* es relativamente alta, la *recall* es considerablemente baja. Esto se debe a que el conjunto de datos se encuentra desbalanceado en términos de la clase a predecir y el modelo de regresión logística no logra encontrar un método adecuado para clasificar.

Tabla 3

Evaluación del modelo de regresión logística sobre la partición de evaluación. Métricas relevantes considerando a "corta duración" como la clase relevante

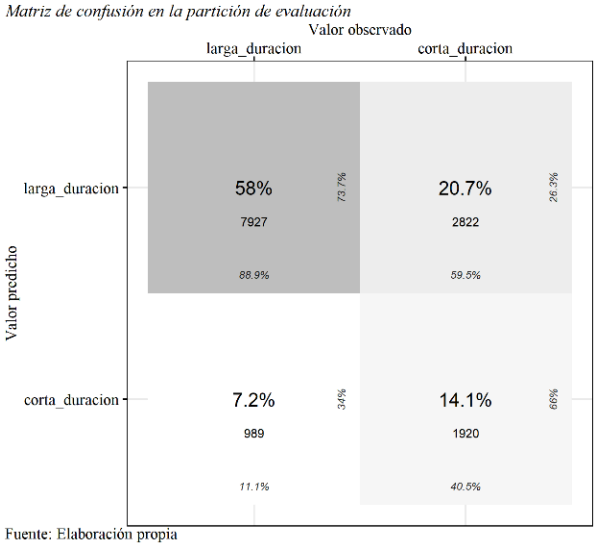
Métrica	Valor
Accuracy	0,72
Recall	0,40
Precisión	0,66
Área bajo la curva ROC	0,74

Fuente: Elaboración propia

Por otro lado, se incluye la matriz de confusión asociada al modelo en la Figura 6. Tal como se observa, la *accuracy* elevada es principalmente por clasificar correctamente a la clase mayoritaria (larga duración). Sin embargo, este modelo no logra clasificar muchas observaciones de la clase minoritaria (corta duración).

Figura 6

Matriz de confusión en la partición de evaluación



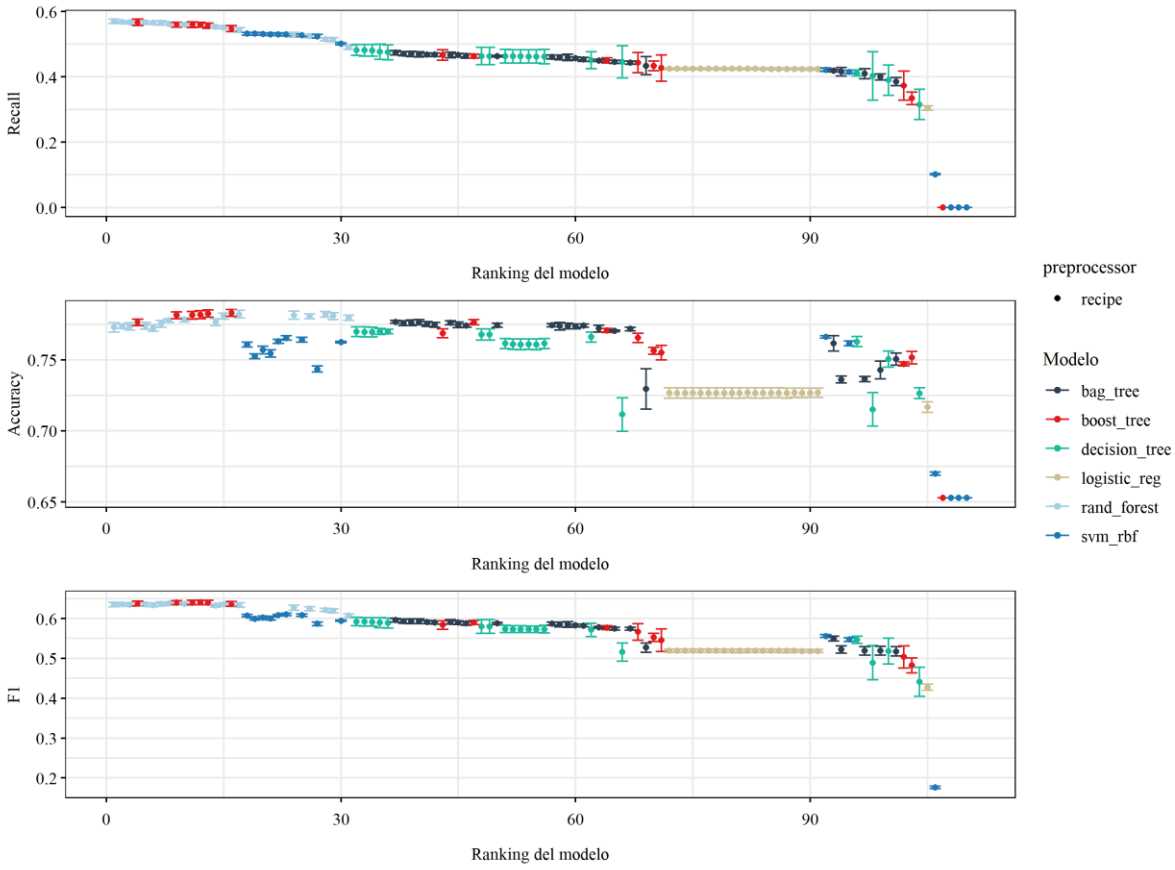
Fuente: Elaboración propia

Considerando a la regresión logística como modelo base, se busca mejorar la clasificación mediante la implementación de modelos de aprendizaje automático. En este sentido, se ajustan los modelos definidos en la sección 2.1. Se utiliza la técnica de optimización bayesiana para el ajuste de hiperparámetros. Esta técnica permite arribar a mejores modelos mediante un ajuste más rápido que la optimización con *grid search*. En este método de optimización, se generan N posibles combinaciones de hiperparámetros para cada uno de los modelos a ajustar. Las combinaciones iniciales son aleatorias y luego de haberlas entrenado, comienza el proceso bayesiano de búsqueda de hiperparámetros en las zonas en donde mejor performa cada iteración. El ajuste se realiza con validación cruzada, con lo cual es posible obtener la métrica promedio de cada iteración para cada modelo, junto con el desvío. En este caso, se realiza la optimización en función de la métrica *recall*, almacenando el resto de las métricas relevantes para la posterior evaluación.

En la Figura 7 se presentan los resultados de la optimización. Se observa que el mejor modelo fue el *random forest* en la métrica *recall* y su desvío fue bajo. Sin embargo, en términos de *accuracy y F1-score* se obtuvieron mejores resultados con el *boosted tree* y, por esta razón, este será el modelo seleccionado ya que su *performance* en la *recall* fue muy similar a la del *random forest*. Es importante destacar que a pesar de que la métrica es baja, es considerablemente superior a un modelo de regresión logística. La variable por predecir (clase) estaba desbalanceada y tener una métrica de este valor es un buen indicador. Además, el resto de las métricas asociadas a este modelo también representan un buen ajuste.

Figura 7

Evaluación de modelos en el ajsute de hiperparámetros mediante optimización bayesiana con cross validation



Fuente: Elaboración propia

En base a los resultados anteriores, se selecciona el mejor modelo en función de la métrica seleccionada. En este caso, se selecciona un modelo *boost tree* con 911 árboles, 86 observaciones como mínimo en cada nodo, profundidad de 4 y tasa de aprendizaje de 0.071. Con esta especificación se vuelve a entrenar el modelo, con todos los datos de entrenamiento ya que anteriormente se había realizado el ajuste con *cross validation.*

Con el nuevo modelo ajustado, se realiza la evaluación en la partición de evaluación. En la Tabla 4 se presentan las métricas evaluadas. Se observa que la clase mayoritaria es “larga duración”, el error de falsos positivos es bajo, es decir, existen relativamente pocas observaciones predichas como “corta duración” que en realidad fueron larga duración.

Tabla 4

Evaluación del modelo de clasificación sobre la partición de evaluación. Métricas relevantes considerando a "corta duración" como la clase relevante

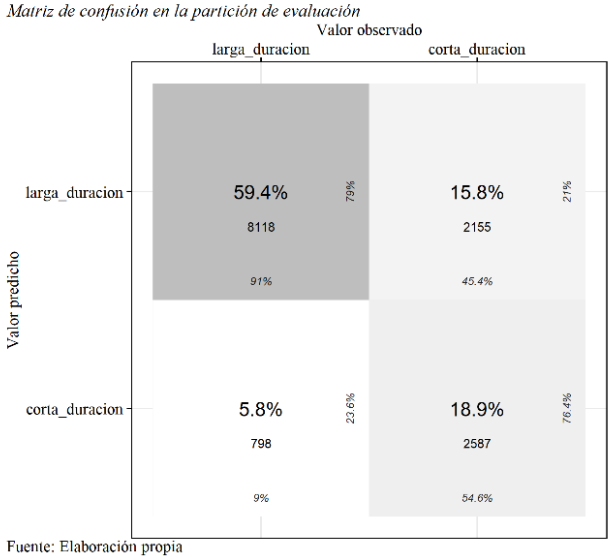
Métrica	Valor
Accuracy	0,78
Recall	0,55
Precisión	0,76
Área bajo la curva ROC	0,84

Fuente: Elaboración propia

En la Figura 8 se incluye la matriz de confusión correspondiente a los datos de evaluación. Hay 1971 observaciones clasificadas erróneamente como “larga duración” que en realidad tuvieron una baja anticipada. Estos registros (falsos negativos) son observaciones que el modelo no logró clasificar adecuadamente. Sin embargo, es necesario destacar que para el caso de estudio este tipo de error es esperable: podría tratarse de bajas de vehículos luego de accidentes de tránsito, lo cual es más difícil de predecir con las variables consideradas.

Figura 8

Matriz de confusión en la partición de evaluación

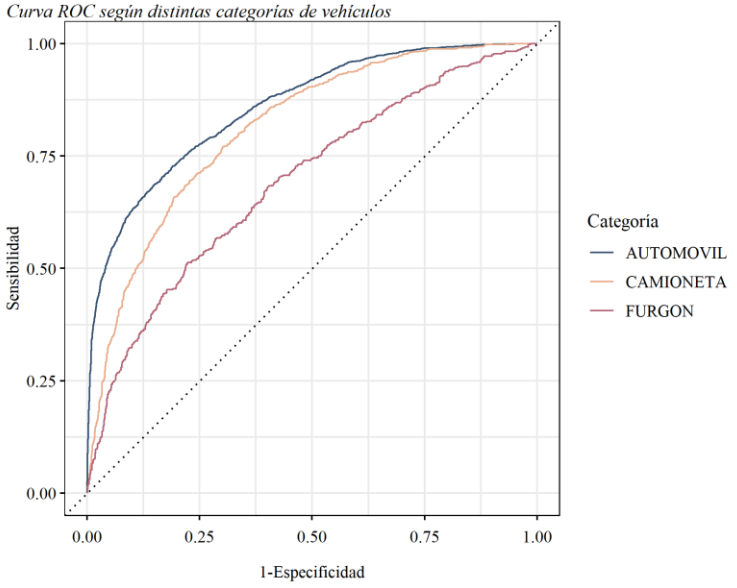


Fuente: Elaboración propia

Hasta aquí se realizó la evaluación del modelo en términos agregados. En la Figura 9 se presenta la curva ROC para las distintas categorías de vehículos. Se observa que el modelo realiza mejores clasificaciones en el caso de automóviles y peores clasificaciones en los vehículos de tipo furgón, siendo las camionetas un caso intermedio.

Figura 9

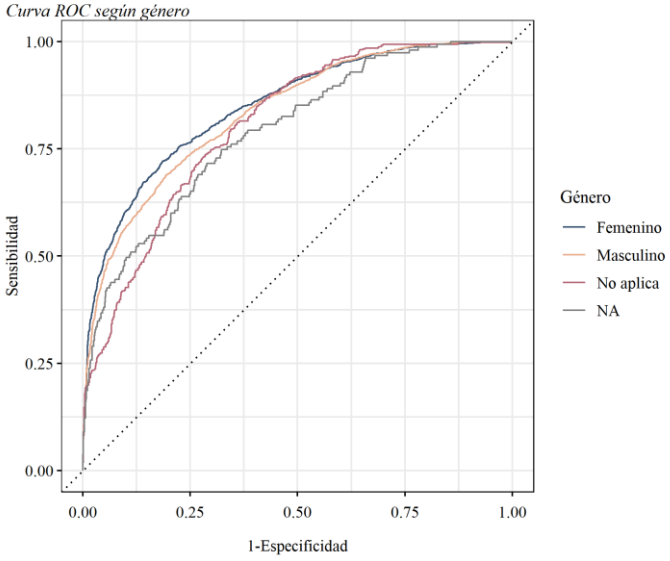
Curva ROC según distintas categorías de vehículos



Considerando el género, en la Figura 10 se observa que el modelo predice mejor para los casos de titulares con género femenino. Existen casos en los cuales el género no aplica dado que se trata de personas jurídicas. En este sentido, el modelo predice mejor para los casos de personas físicas.

Figura 10

Curva ROC según género

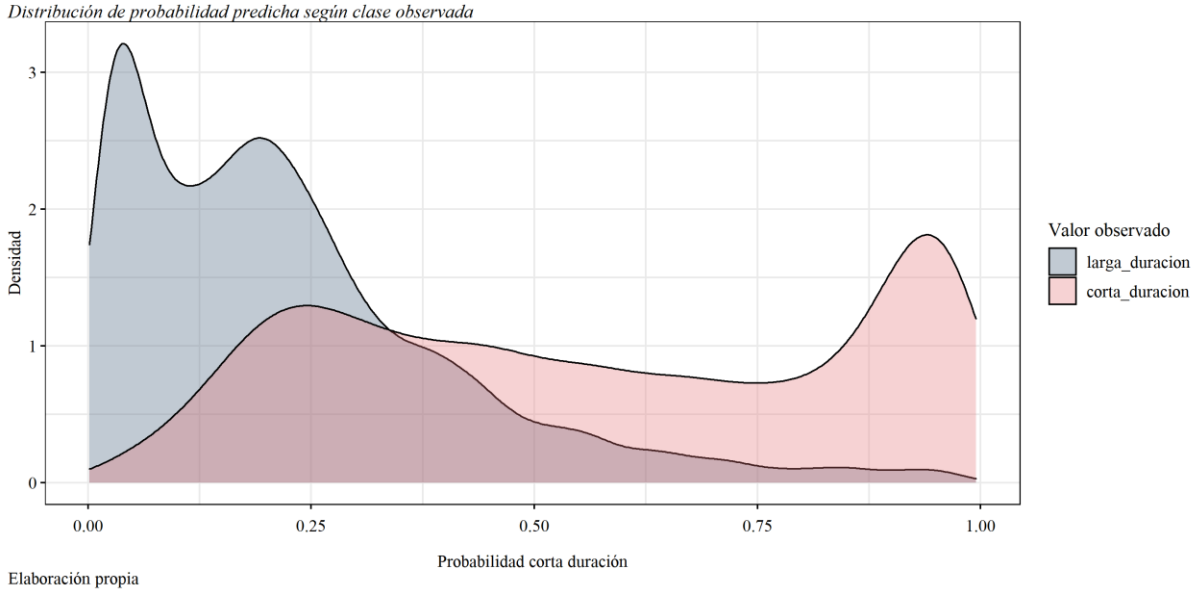


En relación con el valor observado, la Figura 11 muestra la relación entre la probabilidad predicha y el valor real. Se observa que para vehículos que han tenido larga duración en general se estima una baja probabilidad de corta duración. Si se considera el umbral en 0.5, tal como se hizo para la evaluación de las distintas métricas, existen ciertas observaciones en las cuales el modelo predice una baja anticipada y en realidad tuvieron larga duración. Este tipo de error es

relevante en el caso de análisis: son vehículos que, de darse de baja, lo harían en el largo plazo y el modelo predice que será a corto plazo. En la sección 3.2 se analizará este aspecto en detalle. Por otro lado, en relación con los vehículos de corta duración la probabilidad predicha indica que el modelo no logra clasificar bien ciertas observaciones: la probabilidad predicha para ciertas observaciones es considerablemente baja mientras tuvieron corta duración. Esto podría deberse a accidentes o cuestiones más difíciles de predecir.

Figura 11

Distribución de probabilidad predicha según clase observada



Elaboración propia

3. Implementación de modelos para la toma de decisiones

En esta sección se busca transformar el análisis en decisiones. Habiendo definido un modelo para entender el comportamiento de los consumidores que actúan en el mercado analizado, se define un segmento de interés: vehículos con inscripciones recientes. Se generan predicciones para obtener la probabilidad de baja anticipada de los vehículos nacionales incluidos en el segmento definido. Se realiza una ponderación de estos resultados por la proporción de vehículos que se dan de baja en cada modelo y, de esta forma, se obtiene la cantidad de vehículos que se darán de baja anticipadamente. Estos vehículos deben ser tenidos en cuenta por las organizaciones de la cadena autopartista nacional, ya que gran parte de las bajas anticipadas corresponden a bajas con recuperación de partes. Luego de haber trabajado con un gran volumen de datos, se presentan aspectos para mejorar la gestión de los datos correspondientes al mercado automotor.

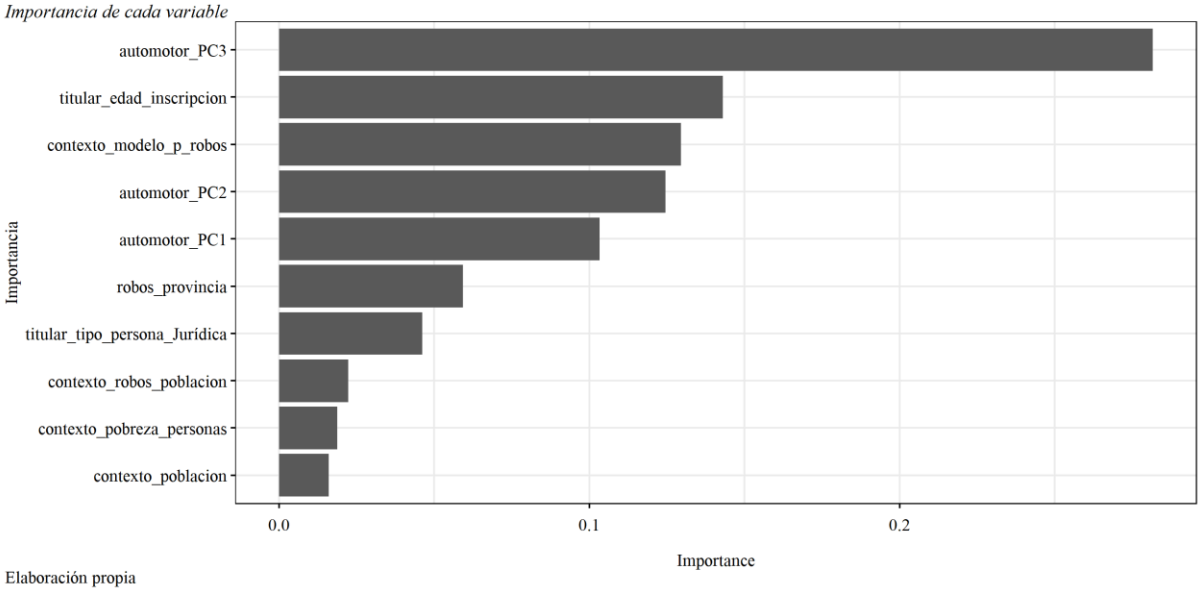
3.1. Aplicación de modelos analíticos predictivos

Luego de seleccionar el mejor modelo, y considerando aceptables las métricas de evaluación en la partición de evaluación, es necesario reentrenar el modelo con los datos completos: la partición de entrenamiento junto con la partición de evaluación. De esta forma, se incorporan los datos adicionales de evaluación con el objetivo de mejorar el ajuste del modelo, dado que ya se conoce la mejor especificación. El nuevo modelo entrenado con todos los datos queda listo para implementarse ya que ha sido entrenado con todos los datos posibles. Sin embargo, ya no es posible evaluarlo debido a que se realizarán predicciones en nuevos datos de inscripciones vehiculares, en donde aún no se conoce la fecha de baja.

En la Figura 12 se presenta la importancia de las variables en este modelo. Se observa que entre las principales 10 variables aparecen las variables correspondientes al análisis de componentes principales realizado en la sección 2.2. Además, también son relevantes aspectos como la edad del titular o aspectos del contexto como el porcentaje de robos, la pobreza o cuestiones demográficas. El tipo de persona también ha sido relevante.

Figura 12

Importancia de cada variable



Elaboración propia

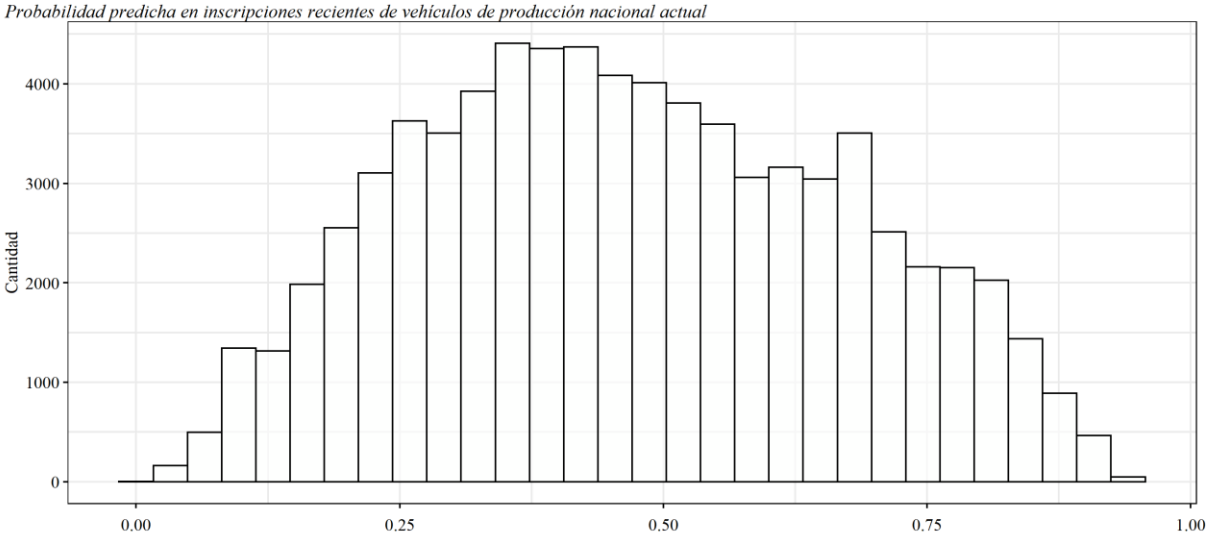
De esta forma, este modelo ya puede ser utilizado para generar inferencias en datos nuevos, de los cuales aún no se conoce la categoría. Es importante destacar que la evaluación de importancia de variables es una forma de interpretación del modelo. Sin embargo, esto no permite cómo afecta cada una de las variables a la probabilidad predicha. Para ello, existen otros métodos, como es el caso de *Shapley Values*, que será analizado sobre las inferencias en nuevas inscripciones al finalizar este apartado.

3.2. Análisis de resultados de las predicciones obtenidas

El objetivo del trabajo es predecir cuántas bajas anticipadas ocurrirán en las nuevas inscripciones realizadas en cada uno de los registros seccionales del país. Para ello, se cuenta con 75119 registros de inscripciones realizadas en el año 2021, de las cuales no se tiene la fecha de baja. Estas inscripciones son de vehículos relevantes para la ADEFA ya que se están produciendo actualmente por las empresas agrupadas en esta organización. Existe la posibilidad de que algunos de estos vehículos ya hayan sido dados de baja pero no se cuenta con un identificador único del vehículo que permita vincular los datos de inscripciones con los datos de bajas. En la Figura 13 se presenta la distribución de probabilidad de baja temprana, si se diera la baja.

Figura 13

Probabilidad predicha en inscripciones recientes de vehículos de producción nacional actual



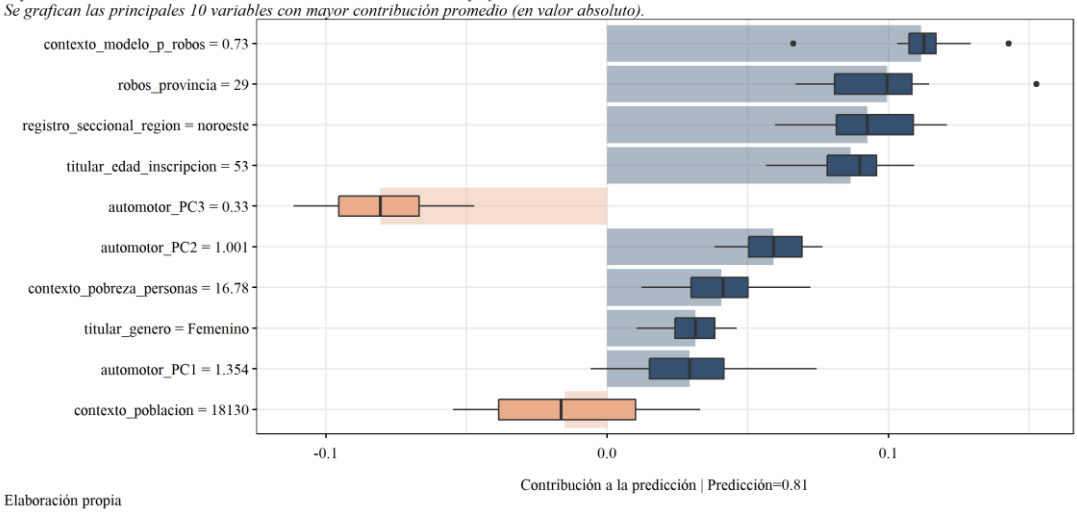
Elaboración propia

En la sección 3.1 se presentó la importancia de las variables en el modelo final. Esto permite entender qué variables han sido relevantes para clasificar a una observación en corta o larga duración. Sin embargo, luego de haber observado la distribución de probabilidad predicha para todas las inscripciones relevantes surge el interrogante: ¿cómo impacta cada variable en cada predicción? Una forma de explicar los modelos denominados "caja negra" (*"black box"*) es mediante el uso de *SHAP values* (*Shapley Additive explanations)*, (Lundberg y Lee, 2017). Este método permite asignar determinada importancia para cada variable en cada predicción.

En las Figuras 14 y 15 se presentan los *shap values* para dos observaciones aleatorias. En Este caso, se observa que la variable más influyente difiere entre ambos casos. En el primer caso, la variable de contexto_modelo_p_robos tenía un valor original de 0.73 e influyó positivamente sobre la probabilidad de baja anticipada. Otras variables que influyeron positivamente fueron la región (Noroeste), la edad (53 años), el componente principal 2 de automotor, la pobreza y el género (femenino). El componente principal 3 correspondiente al modelo de vehículo y la cantidad de población de la zona influyeron negativamente pero el efecto del resto de las variables fue superador y, por esta razón, la probabilidad predicha es 0.81: corta duración.

Figura 14

Explicación de una observación aleatoria mediante el método de shapley values.
Se grafican las principales 10 variables con mayor contribución promedio (en valor absoluto).

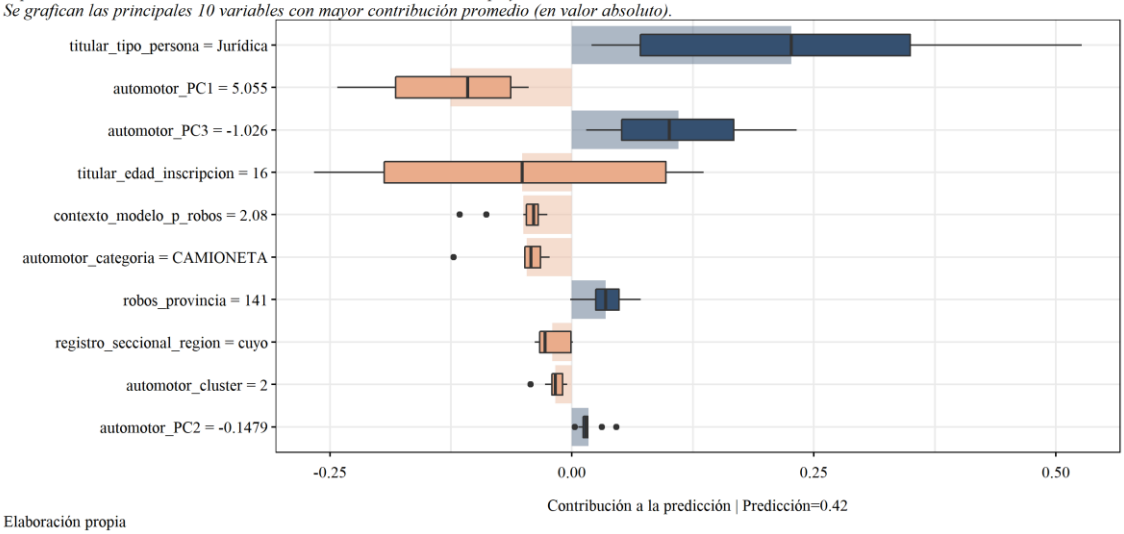


Elaboración propia

En el caso de la observación presentada en la Figura 15, las variables que más influyeron son distintas al caso anterior: el tipo de persona es una variable que influyó considerablemente a incrementar su probabilidad de baja anticipada. Sin embargo, otros aspectos tuvieron influencia, haciendo que la probabilidad final sea 0.42: larga duración. Notar que en este caso los componentes principales influyen en forma inversa al caso anterior, indicando que se trata de un vehículo considerablemente diferente al anterior.

Figura 15

Explicación de una observación aleatoria mediante el método de shapley values.
Se grafican las principales 10 variables con mayor contribución promedio (en valor absoluto).



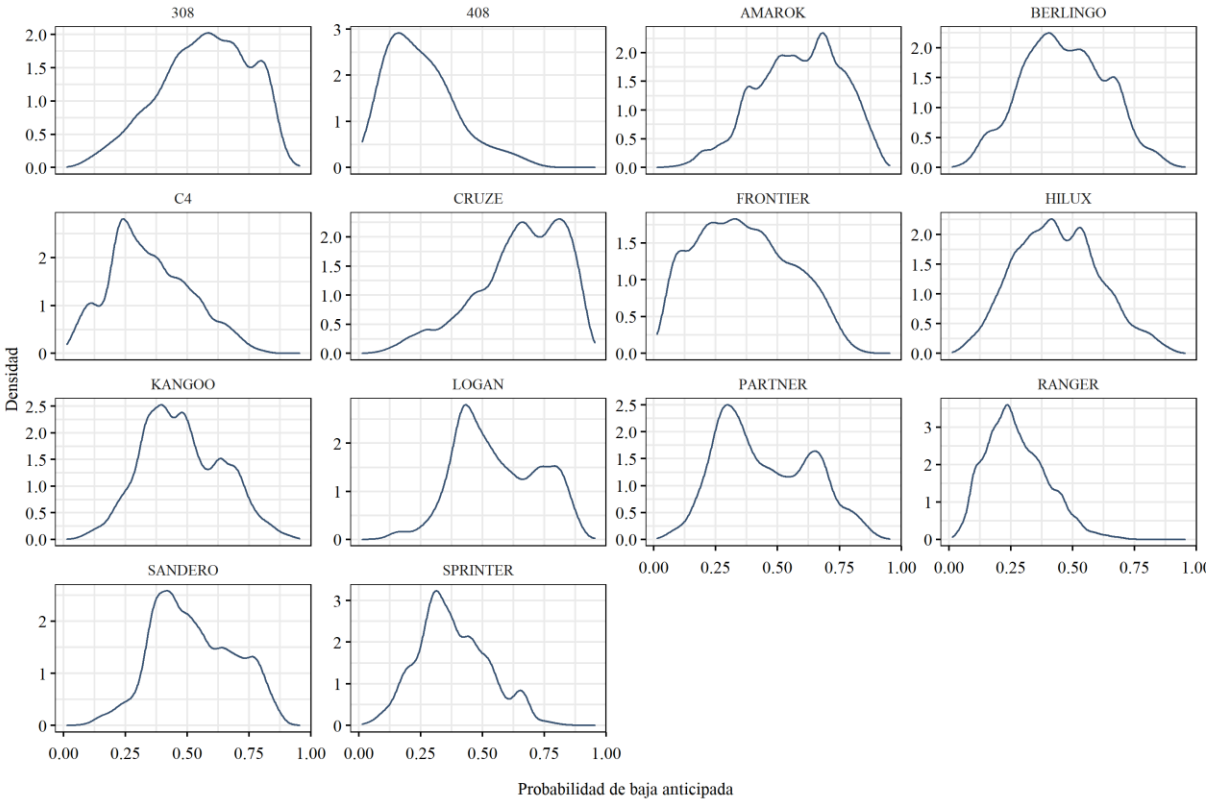
Elaboración propia

En el contexto organizacional de la organización ADEFA, se espera mejorar la toma de decisiones mediante el uso de datos y métodos analíticos. En este caso, se obtuvieron predicciones de la probabilidad de baja anticipada para cada uno de los vehículos inscriptos durante el año 2021. En la Figura 16 se presenta la distribución de probabilidad predicha para cada uno de los modelos de vehículos relevantes inscriptos. Se observa que el modelo predice

probabilidades de baja anticipada menores para ciertos vehículos y mayores para otros. Sin embargo, el análisis es a nivel de vehículo individual, en donde además de las características del vehículo existen aspectos del titular o contexto que influyen sobre esta probabilidad. Las diferencias en probabilidad en vehículos del mismo modelo se deben a este tipo de cuestiones.

Figura 16

Distribución de probabilidad estimada de baja anticipada según modelo de automotor



Tal como se ha mencionado, las predicciones representan la probabilidad de baja temprana, si se diera la baja. Al comienzo del trabajo se visualizó la evolución de las inscripciones y bajas y la cantidad de vehículos inscriptos en cada momento es muy grande. Parte de esta diferencia es un problema temporal: a medida que pasa el tiempo se inscriben más vehículos con el crecimiento poblacional y las bajas están rezagadas. Sin embargo, las diferencias observadas son de una magnitud muy grande y permiten pensar que no es solo un problema de rezago, sino que existen ciertos vehículos que no son dados de baja.

Para poder retornar una probabilidad de baja anticipada para cada una de las inscripciones, es necesario ponderar la probabilidad obtenida por la probabilidad de baja. Dado que sólo se cuenta con datos de bajas e inscripciones en cierto período, estimar esta probabilidad es complicado. De momento, se considerará el porcentaje de bajas sobre inscripciones de cada

uno de los modelos. Se entiende que esto podría mejorarse al disponer de los datos faltantes, pero sirve como aproximación para poder presentar resultados más adecuados. En la Tabla 15 se presenta la cantidad de inscripciones y bajas por modelo de vehículo durante el período 2018-2021. Se incluye una columna correspondiente al porcentaje de bajas. Además, en función de la probabilidad predicha para cada vehículo inscripto en 2021 y un umbral de 0.5, se obtiene el porcentaje de vehículos de corta duración, Entonces, considerando el porcentaje de bajas por vehículo y el porcentaje de vehículos con corta duración, se añaden las cantidades de bajas anticipadas esperadas por modelo de vehículo.

Tabla 5

Predicciones según modelo de vehículo

Modelo	Bajas	Inscripciones	% Bajas	Inscripciones 2021	% Corta duración	Bajas anticipadas
KANGOO	2998	39437	7.60%	6786	41.23%	213
SANDERO	2783	49317	5.64%	6451	50.92%	185
AMAROK	818	58663	1.39%	14883	69.20%	144
PARTNER	2038	24160	8.44%	3952	37.58%	125
LOGAN	1426	29458	4.84%	3475	55.86%	94
CRUZE	521	32682	1.59%	6557	80.36%	84
308	1014	12477	8.13%	1470	65.58%	78
HILUX	2184	84656	2.58%	7378	36.26%	69
BERLINGO	1000	12002	8.33%	1869	41.79%	65
C4	1358	25394	5.35%	5114	18.75%	51
SPRINTER	336	10689	3.14%	3118	20.88%	20
RANGER	1680	51618	3.25%	11832	4.49%	17
FRONTIER	136	6602	2.06%	2168	25.74%	11
408	734	4714	15.57%	66	6.06%	1

Fuente: Elaboración propia en base a datos de la DNRPA

3.3. Recomendaciones para una gestión eficiente de datos

El análisis de datos en contextos de grandes volúmenes provenientes de diversas fuentes es complejo. Existen diversos aspectos que deben ser tenidos en cuenta al momento de consolidar este tipo de datos. En primer lugar, el mantenimiento de un esquema de análisis ordenado es esencial. En este trabajo se ha realizado un gran esfuerzo para generar un conjunto de datos adecuado para la aplicación de modelos.

Se presentarán una serie de cuestiones para tener en cuenta con el objetivo de mejorar los datos utilizados. Más allá de estos aspectos, se destaca el rol de la DNRPA en la disponibilización de estos datos ya que son de gran relevancia y la publicación se realiza mensualmente hasta la actualidad. Los datos abiertos actúan como un bien público que permite mejorar la asignación de recursos desde el sector público-privado mediante el entendimiento de las dinámicas de los comportamientos individuales.

En primer lugar, se considera necesaria la adecuación de los datos disponibilizados desde la DNRPA que incluya un identificador único de vehículo para poder realizar cruces entre los datos existentes: inscripciones, transferencias, robos y bajas. Esto permitiría hacer un seguimiento de todo lo ocurrido a nivel vehículo, obteniendo *insights* más apropiados.

En relación con las variables categóricas, existen una gran cantidad de categorías únicas que al parecer se están registrando en formato de texto libre. Este es el caso de los campos de localidad, provincia, país, marca y modelo de automotor. No contar con identificadores únicos de cada categoría implica la necesidad de grandes esfuerzos para consolidar categorías y, a su vez, esto puede generar errores de agrupamientos. En este sentido, el método de registro de este tipo de campos debería ser mediante campos de selección entre una lista desplegable de categorías previamente identificadas.

Otro problema asociado a la escritura manual de categorías es que se han encontrado categorías erróneamente imputadas. Por ejemplo, se han encontrado vehículos de un modelo registrados en una marca que no corresponde. Este es un gran problema dado que, en el contexto de grandes volúmenes de datos, identificar este tipo de errores es muy complejo. Se ha realizado un trabajo detallado de identificación de categorías erróneas, pero resulta difícil determinar en donde existen este tipo de problemas de registro.

Se cuenta con un conjunto de datos con información de cada uno de los registros seccionales que incluye la dirección, pero no se encuentra normalizada. Sería de gran utilidad incorporar a estos datos la latitud y longitud de cada registro en vez de la dirección ya que esto permitiría trabajar los datos en forma más simple incorporando análisis espaciales considerando la distancia hasta el registro.

Conclusión

En este trabajo se ha desarrollado una metodología de análisis de datos provenientes de la DNRPA. No existen muchos estudios realizados con estos datos y son de gran relevancia dado el vínculo estrecho que mantienen con uno de los principales sectores productivos a nivel nacional: la industria automotriz. En este sentido, se realizó un estudio comenzando por la extracción y limpieza de datos, los cuales fueron enriquecidos mediante datos adicionales vinculados a aspectos del contexto argentino. Mediante el análisis exploratorio de los datos, se encontró una distribución bimodal de la duración de la tenencia y esto permitió generar una clasificación de las observaciones: vehículos de corta duración y vehículos de larga duración.

Mediante la utilización de modelos estadísticos multivariados y de aprendizaje automático, se estableció un modelo para estimar la probabilidad de una baja anticipada de un vehículo. El modelo que mejor realiza la clasificación en larga y corta duración fue utilizado para clasificar a las nuevas inscripciones. Esta clasificación corresponde a la probabilidad de que la baja, de ocurrir, lo haga en el corto plazo. Hay dos enfoques posibles para utilizar el resultado de las predicciones: ponderando el porcentaje de cada clase en cada modelo de vehículo por la proporción de bajas que se han realizado en ese modelo de vehículo o la otra alternativa es utilizar la predicción de probabilidad como un *score* de riesgo de baja anticipada.

Originalmente se esperaba establecer un modelo para predecir la duración de vehículos en años, es decir: un modelo de regresión. Dado que se cuenta con datos restringidos a cierto período, en donde la duración sigue una distribución bimodal, la transformación del problema en uno de clasificación permitió clasificar a las observaciones con cierta confianza. Mediante métodos de explicación de modelos de aprendizaje automático fue posible realizar una interpretación de los resultados, encontrando cuáles fueron las variables más relevantes para clasificar a los vehículos en corta y larga duración y, asimismo, comprender en qué sentido operan los efectos para observaciones individuales.

Luego de haber seleccionado el mejor modelo, se propuso un método para obtener la cantidad de bajas anticipadas esperadas por modelo de vehículo. En este sentido, se realizaron predicciones para todas las inscripciones de vehículos de producción nacional actual inscriptos durante el año 2021. La probabilidad de baja anticipada obtenida fue ponderada por la proporción de bajas que se registran en cada vehículo.

El trabajo se centró en analizar la situación desde el enfoque de modelos de clasificación, utilizando el análisis de componentes principales y técnicas de *clustering* como métodos para ingeniería de características. Sin embargo, dado el tipo de problema, utilizar modelos de supervivencia aplicados a la duración de los vehículos podría ser un enfoque más adecuado, considerando la censura en las nuevas observaciones, donde aún no se sabe en qué momento se darán de baja. Para ello, sería de gran utilidad contar con un intervalo más amplio de los datos. Además, si se contara con datos de un intervalo más grande también podría trabajarse de mejor forma con modelos de regresión para estimar la duración de los vehículos.

El análisis se focalizó en los vehículos de producción nacional en el año 2021, con énfasis en mejorar la toma de decisiones por parte de la ADEFA. Al haber trabajado con datos de la DNRPA, se han mencionado una serie de oportunidades para la mejora en la gestión de datos para maximizar la utilidad de su uso. No se ha mencionado aspectos relacionados al sector asegurador, pero los resultados podrían ser de relevancia para futuros análisis en este rubro. Además, al tratarse de datos de inscripciones y bajas a lo largo del tiempo, podría trabajarse en modelos de series de tiempo para la gestión de la demanda de trámites en la DNRPA.

Como comentario final, es necesario destacar que este trabajo se ha realizado haciendo énfasis en el uso de los métodos para la gestión de datos en contextos organizacionales. Un análisis más detallado del contexto del mercado automotor permitiría mejorar los métodos y obtener resultados más concluyentes. Sin embargo, al haber realizado el procesamiento completo de los datos, evaluación y posterior implementación, se cuenta con una base para futuros análisis.

Referencias bibliográficas

Anowar, S., Eluru, N., & Miranda-Moreno, L. (2014). Alternative modeling approaches used for examining automobile ownership: a comprehensive review. *Transport Reviews, 34(4)*, 441-473.

Banco Mundial. (Noviembre, 2021). *Transporte*. Entendiendo la pobreza: https://www.bancomundial.org/es/topic/transport/overview

Breiman, L. (1996). Bagging predictors. *Mach Learn 24*, 123–140.

Breiman, L. (2001). Random forests. *Machine learning, 45(1)*, 5-32.

Breiman, L., Friedman, J., Olshen, R., & Stone , C. (2017). Classification and regression trees. *Routledge*.

Canatrella, J., Guzmán, G., & Katz, L. (2008). La industria automotriz argentina: limitantes a la integración local de autocomponentes. *Documento Técnico, 1*.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM transactions on intelligent systems and technology (TIST), 2(3)*, 1-27.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *n Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785-794.

Clark, B. C. (2016). Changes in level of household car ownership: the role of life events and spatial context. *Transportation, 43(4)*, 565-599.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software, 33(1)*, 1–22.

Gentile, G., & & Nökel, K. (2016). *Modelling public transport passenger flows in the era of intelligent transport systems.* Springer.

González, F. A. (2004). Teoría de la decisión e incertidumbre: modelos normativos y descriptivos. *Empiria. Revista de metodología de ciencias sociales, (8)*, 139-160.

Guo, Z. (2013). Does residential parking supply affect household car ownership? The case of New York City. *Journal of Transport Geography, 26*, 18-28.

Hall, P., & York, M. (2001). On the calibration of Silverman test for multimodality. 515–536.

Joyanes, A. L. (2016). *Big Data, Análisis de grandes volúmenes de datos en organizaciones.* Alfaomega Grupo Editor.

Kuhn, M., & Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning*. https://www.tidymodels.org.

Lee, W. T., & Govindan, S. (2014). Emerging issues in car purchasing decision. *Academic Research International, 5(5)*, 169.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information*, 4768-4777.

Mergel, I., Rethemeyer, R., & Isett, K. (2016). Big data in public affairs. .*Public Administration Review, 76(6)*, 928-937.

Moody, J., Farr, E., Papagelis, M., & Keith, R. D. (2021). The value of car ownership and use in the United States. *Nature Sustainability*.

Müller, A. (2013). El transporte en la Región Metropolitana de Buenos Aires ¿hacia el "colapso"? *Documento de Trabajo del Centro de Estudios de la Situación y Perspectivas de la Argentina*.

Polyzotis, N., Roy, S., Whang, S., & Zinkevich. (2018). Data lifecycle challenges in production machine learning: a survey. *ACM SIGMOD Record, 47(2),*, 17-28.

Sayyadi, R., & Awasthi, A. (2017). A system dynamics based simulation model to evaluate regulatory policies for sustainable transportation planning. *International Journal of Modelling and Simulation, 37(1)*, 25-35.

Tsamboulas, D., Verma, A., & Moraiti, P. (2013). Transport infrastructure provision and operations: Why should governments choose private–public partnership? *Research in Transportation Economics, 38(1)*, 122-127.

Yamamoto, T. (2008). The impact of life-course events on vehicle ownership dynamics: The cases of france and Japan. *IATSS research, 32(2)*, 34-43.

Apéndice A: Organizaciones

En esta sección incluye recursos que complementan el análisis de las organizaciones relevantes y el contexto en el cual se desenvuelven.

Tabla A1

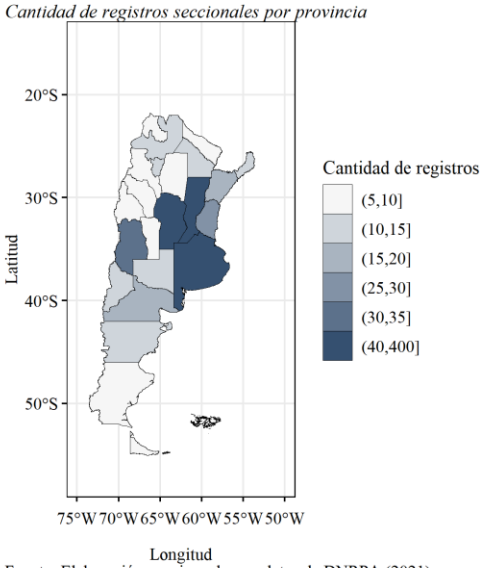
Empresas agrupadas en la Asociación de Fabricas Automotores (ADEFA)

	Empresa	Ubicación	Vehículos
FCA	FCA Automobiles Argentina S.A.	Córdoba	Cronos
Ford	Ford Argentina S.C.A.	Buenos Aires	Ranger
GENERAL MOTORS	General Motors de Argentina S.R.L.	Santa Fé	Cruze
HONDA	Honda Motor de Argentina S.A.	Buenos Aires	Hrv
IVECO	Iveco Argentina S.A.	Córdoba	Tector, Cursor, Stralis, Trakker. Chases Para Buses Eurocargo
	Mercedes-Benz Argentina S.A.	Buenos Aires	Sprinter, Accelo, Atron (1735), Atego (1721 Y 1726) Y Chasis De Buses (Of 1621, Of 1721, Oh 1621, Oh 1721)
NISSAN	Nissan Argentina S.A.	Buenos Aires	Frontier
PSA	PSA Peugeot Citroën Argentina S.A.	Buenos Aires	Peugeot 308, Peugeot 408 Y Citroen C4 Lounge. Peugeot Partner Y Citroen Berlingo
RENAULT	Renault Argentina S.A.	Córdoba	Logan, Sandero, Sandero Stepway, Y Kangoo
SCANIA	Scania Argentina S.A.	Tucumán	Componentes De Transmisión
TOYOTA	Toyota Argentina S.A.	Buenos Aires	Hilux Y Sw4
VOLKSWAGEN	Volkswagen Argentina S.A.	Buenos Aires y Córdoba	Amarok, Componentes

Fuente: Elaboración propia en base a datos de ADEFA, 2021

Figura A1

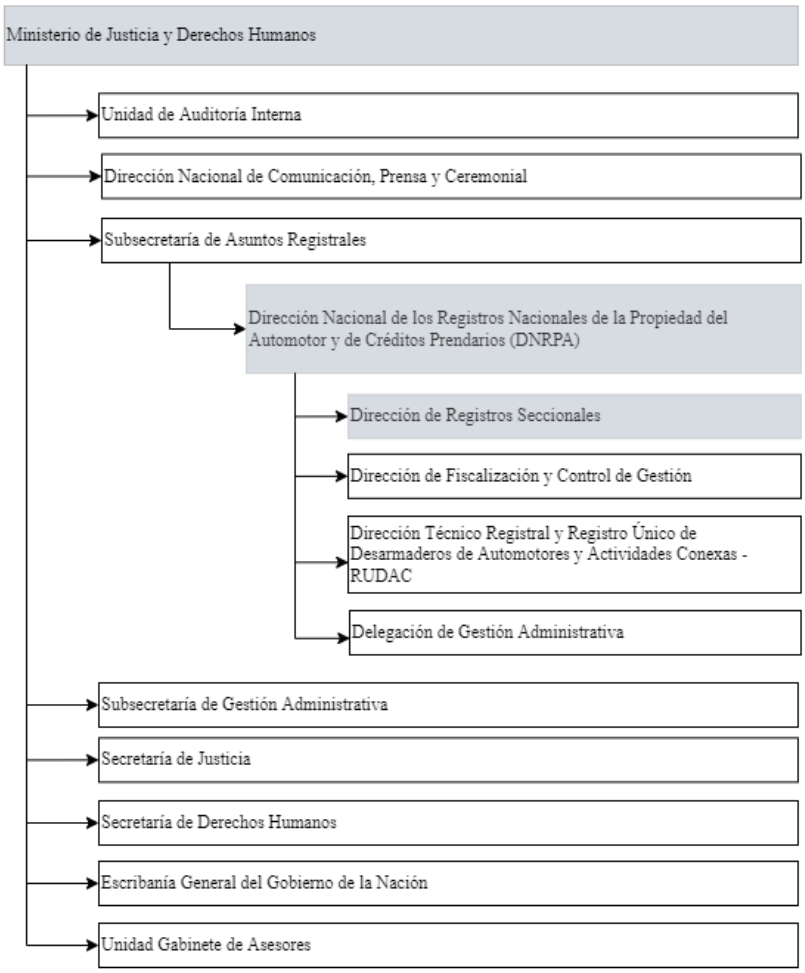
Cantidad de registros seccionales por provincia



Fuente: Elaboración propia en base a datos de DNRPA (2021)

Diagrama A1

Organigrama del Ministerio de Justicia y Derechos Humanos



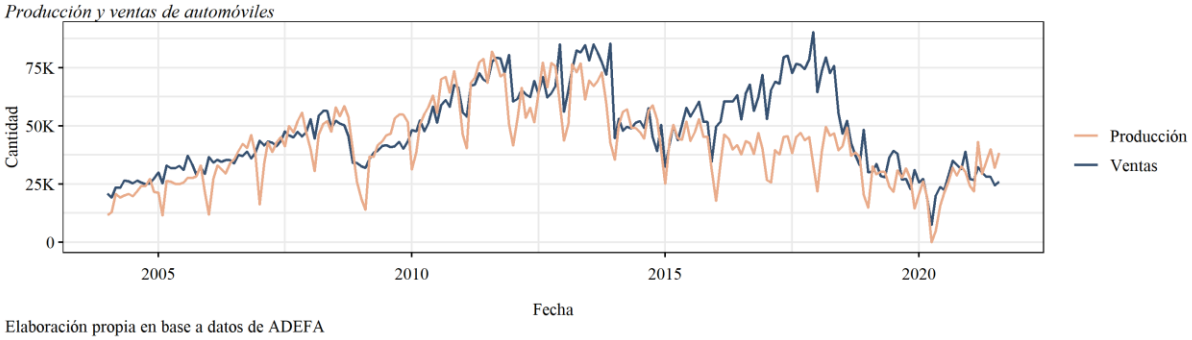
Fuente: Elaboración propia en base a datos del Ministerio de Justicia y Derechos Humanos de la Argentina (2021)

Apéndice B: Estado de situación del mercado automotor argentino

Se incluyen ciertos recursos que complementan el análisis del estado de situación del mercado automotor argentino, considerando la oferta y la demanda de vehículos.

Figura B1

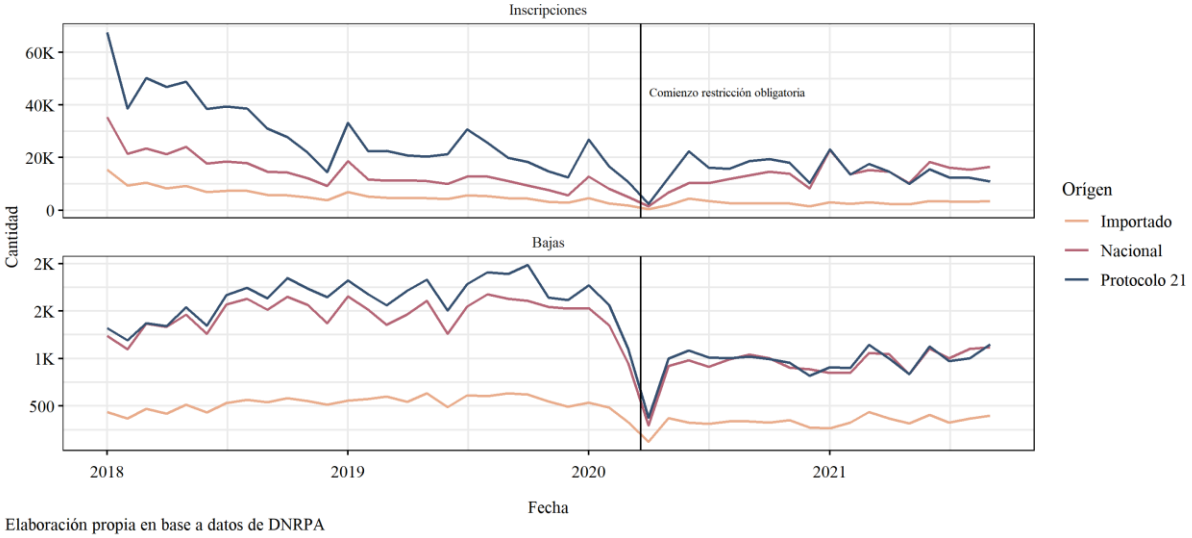
Producción y ventas de automóviles



Elaboración propia en base a datos de ADEFA

Figura B2

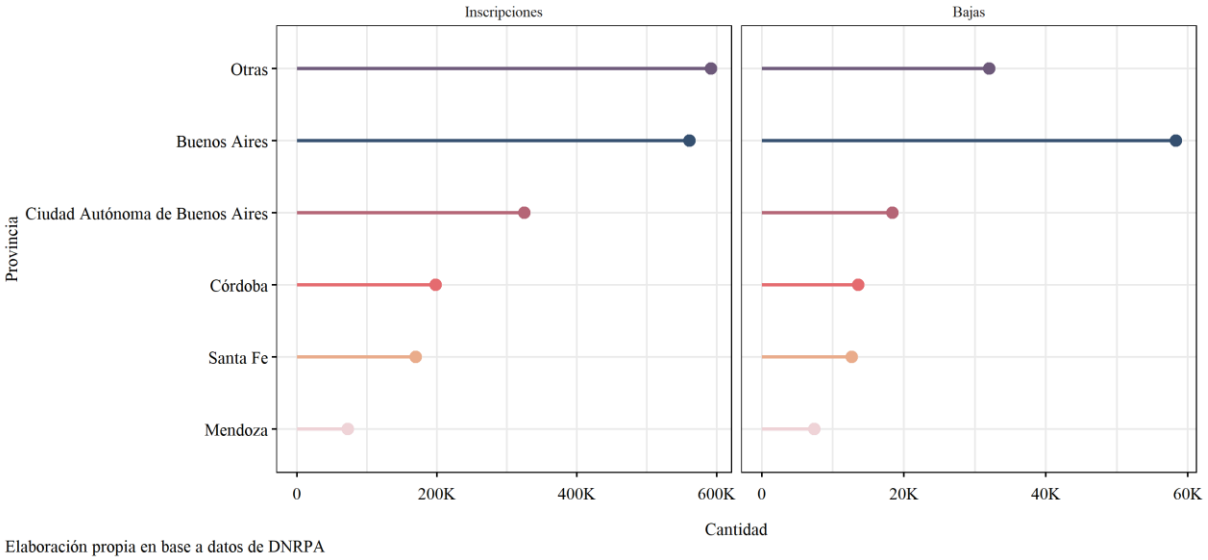
Inscripciones y bajas de vehículos privados en registros seccionales del automotor Argentina 2018-2021



Elaboración propia en base a datos de DNRPA

Figura B3

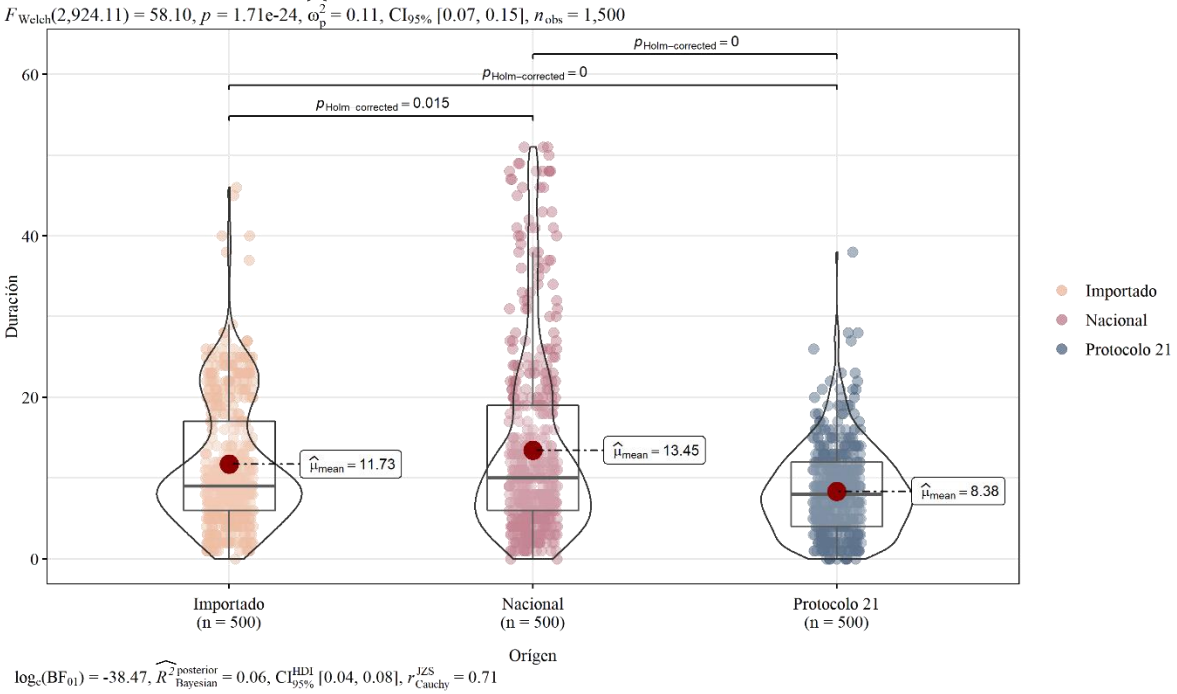
Cantidad de inscripciones y bajas de automóviles privados por provincia



Elaboración propia en base a datos de DNRPA

Figura B4

Diferencias en duración según orígen de vehículos



$\log_e(BF_{01}) = -38.47$, $\widehat{R^2}^{posterior}_{Bayesian} = 0.06$, $CI^{HDI}_{95\%}$ [0.04, 0.08], $r^{IZS}_{Cauchy} = 0.71$

Pairwise test: **Games-Howell test**; Comparisons shown: **only significant**

Se consideró una muestra aleatoria de 500 vehículos estratificada según origen.
Fuente: elaboración propia en base a datos de DNRPA

Apéndice C: Modelo de clasificación de la duración

En esta sección se presentan ciertos aspectos sobre el modelo de clasificación de la duración en corta y larga duración.

Tabla C1

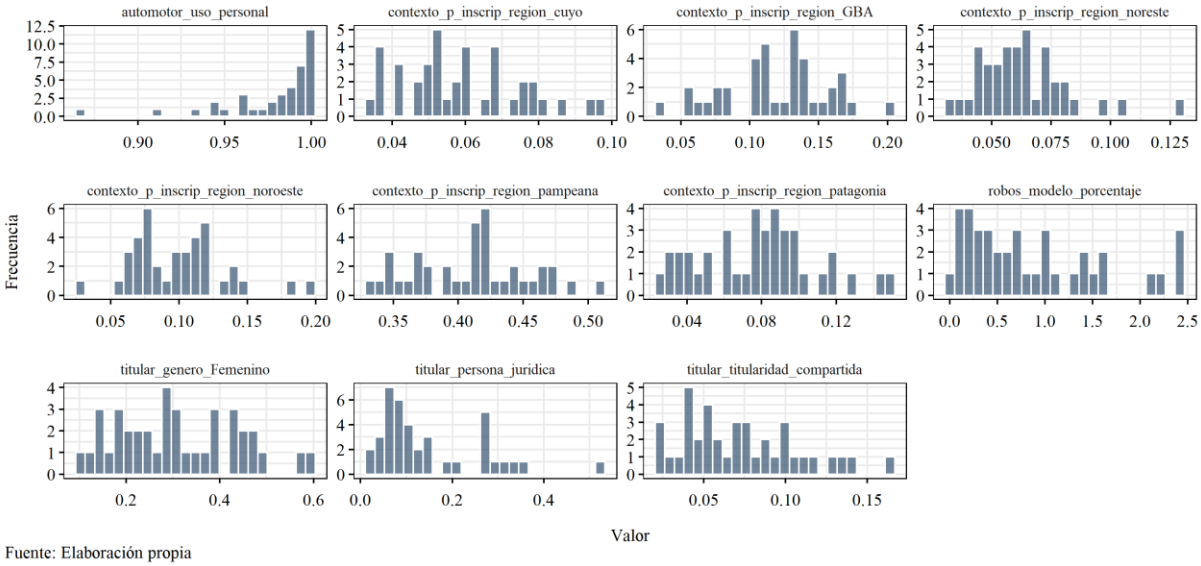
Descripción de las variables utilizadas

Categoría	Variable	Rol	Descripción
	id	id	Identificador único de la observación
	clase	clase	1 = Con baja temprana, 0 = Sin baja temprana
Automotor	automotor_marca	predictor	Marca del vehículo
Automotor	automotor_categoria	predictor	Categoría del vehículo
Automotor	automotor_tipo	predictor	Tipo del vehículo
Automotor	automotor_modelo	predictor	Modelo del vehículo
Automotor	automotor_uso	predictor	Uso del vehículo
Titular	titular_edad_inscripcion	predictor	Edad del titular al momento de la inscripción
Titular	titular_extranjero	predictor	1 = Extranjero, 0 = Argentino
Titular	titular_provincia	predictor	Provincia del titular
Titular	titular_porcentaje_titularidad	predictor	Porcentaje de titularidad del vehículo
Titular	titular_titularidad_compartida	predictor	1 = Titularidad compartida, 0 = Titular único
Titular	titular_genero	predictor	Género del titular: femenino, masculino, no aplica
Titular	titular_tipo_persona	predictor	Tipo de persona: física, jurídica
Ubicación	registro_seccional_provincia	predictor	Provincia del registro seccional
Ubicación	registro_seccional_departamento	predictor	Departamento del registro seccional
Contexto según ubicación	contexto_poblacion	predictor	Población del departamento
Contexto según ubicación	contexto_poblacion_m2	predictor	Población por metro cuadrado del departamento
Contexto según ubicación	contexto_habitantes_hogar	predictor	Cantidad promedio de individuos por hogar en el departamento
Contexto según ubicación	contexto_robos	predictor	Cantidad de robos de vehículos en el departamento
Contexto según ubicación	contexto_robos_poblacion	predictor	Cantidad de robos de vehículos sobre la población del departamento
Contexto según modelo	contexto_modelo_robos	predictor	Cantidad de robos del modelo de vehículo

Fuente: Elaboración propia en base a datos de la DNRPA, INDEC, CEDLAS, 2021

Figura C1

Distribución de las variables a considerar para PCA de modelos de vehículos



Fuente: Elaboración propia

Tabla C2

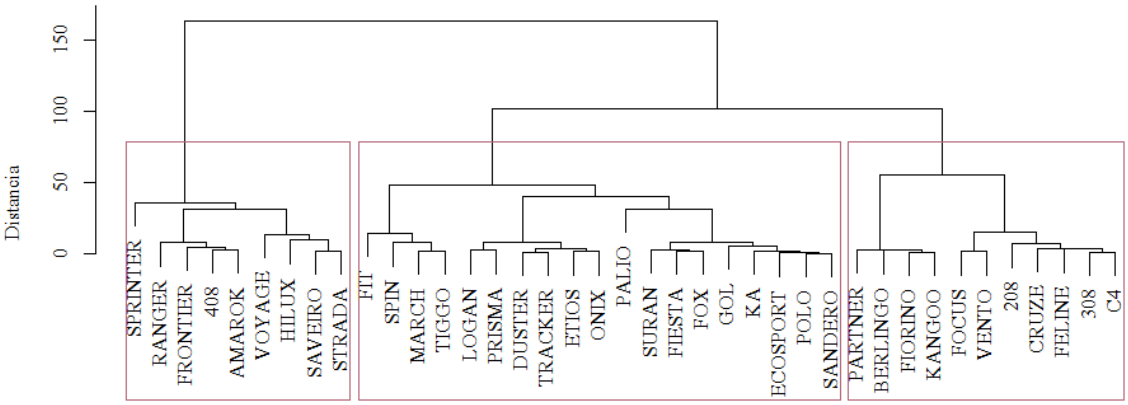
Variabilidad explicada por cada componente

PC	Variabilidad explciada	Variabilidad explicada acumulada
1	32.24%	32.24%
2	18.49%	50.73%
3	13.66%	64.39%
4	12.20%	76.59%

Fuente: Elaboración propia

Figura C2

Dendrograma método de Ward



Fuente: Elaboración propia

Tabla C3

Variables vinculadas al titular: Clasificación según categoría

Característica	larga_duracion, N = 29,718[1]	corta_duracion, N = 15,806[1]	p-valor[2]
titular_extranjero	1,424 (4.8%)	657 (4.2%)	0.002
titular_titularidad_compartida	1,937 (6.5%)	806 (5.1%)	<0.001
titular_genero			<0.001
Femenino	7,662 (31%)	5,157 (34%)	
Masculino	14,667 (59%)	8,560 (56%)	
No aplica	2,477 (10.0%)	1,590 (10%)	
Desconocido	4,912	499	
titular_edad_inscripcion	34 (25, 46)	39 (29, 52)	<0.001
Desconocido	3,114	21	
titular_tipo_persona			<0.001
Física	27,241 (92%)	14,216 (90%)	
Jurídica	2,477 (8.3%)	1,590 (10%)	

[1] n (%); Mediana (RIQ)

[2] prueba chi cuadrado de independencia; prueba de los rangos con signo de Wilcoxon

Fuente: Elaboración propia en base a datos de la DNRPA, 2021

Tabla C4

Variables vinculadas al titular: Clasificación según tipo de persona

Característica	Física, N = 41,457[1]	Jurídica, N = 4,067[1]	p-valor[2]
clase			<0.001
larga_duracion	27,241 (66%)	2,477 (61%)	
corta_duracion	14,216 (34%)	1,590 (39%)	
titular_edad_inscripcion	37 (27, 48)	18 (8, 45)	<0.001
Desconocido	2,887	248	

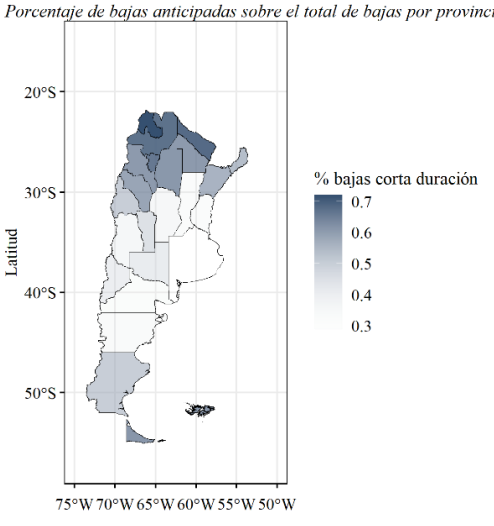
[1] n (%); Mediana (RIQ)

[2] prueba chi cuadrado de independencia; prueba de los rangos con signo de Wilcoxon

Fuente: Elaboración propia en base a datos de la DNRPA, 2021

Figura C3

Porcentaje de bajas anticipadas sobre el total de bajas por provincia



Fuente: Elaboración propia en base a datos de DNRPA.

Tabla C5

Clasificación según región

registro_seccional_region	% Bajas de corta duración
noroeste	65.95%
noreste	58.28%
patagonia	39.80%
cuyo	38.69%
CABA	34.06%
pampeana	30.20%
GBA	27.40%

Fuente: Elaboración propia en base a datos de la DNRPA, 2021

Tabla C6

Variables vinculadas al titular: Clasificación según categoría

Característica	**larga_duracion**, N = 29,718[1]	**corta_duracion**, N = 15,806[1]	**p-valor**[2]
contexto_poblacion	322,375 (115,041, 654,324)	301,572 (107,630, 618,989)	<0.001
contexto_robos_poblacion	0.004 (0.001, 0.006)	0.002 (0.000, 0.006)	<0.001
contexto_pobreza_personas	6.3 (3.7, 9.3)	6.5 (3.7, 9.8)	<0.001
contexto_modelo_p_robos	1.46 (0.96, 3.03)	0.78 (0.43, 2.08)	<0.001

[1] Mediana (RIQ)

[2] prueba de los rangos con signo de Wilcoxon

Fuente: Elaboración propia en base a datos de la DNRPA, 2021

Tabla C7

Resúmen modelos regresión logística

	Automotor	Automotor + titular	Automotor + titular + contexto
(Intercept)	1.411 (0.068)***	1.187 (0.070)***	1.245 (0.087)***
automotor_PC1	0.356 (0.022)***	0.367 (0.022)***	0.424 (0.023)***
automotor_PC2	-0.580 (0.026)***	-0.553 (0.027)***	-0.398 (0.028)***
automotor_PC3	-0.855 (0.018)***	-0.838 (0.018)***	-0.703 (0.020)***
robos_provincia	-0.278 (0.013)***	-0.272 (0.013)***	-0.180 (0.020)***
automotor_categoria_AUTOMOVIL	-1.706 (0.068)***	-1.622 (0.069)***	-1.386 (0.070)***
automotor_categoria_CAMIONETA	-2.087 (0.081)***	-2.040 (0.082)***	-1.875 (0.082)***
automotor_categoria_FURGON			
automotor_cluster_X1	-1.235 (0.059)***	-1.147 (0.060)***	-0.998 (0.061)***
automotor_cluster_X2	-2.107 (0.077)***	-2.045 (0.077)***	-2.042 (0.077)***
automotor_cluster_X3			
titular_edad_inscripcion		0.271 (0.013)***	0.258 (0.013)***
titular_titularidad_compartida		-0.077 (0.013)***	-0.066 (0.013)***
titular_genero_Femenino		0.330 (0.029)***	0.324 (0.029)***
titular_tipo_persona_Jurídica		0.282 (0.048)***	0.390 (0.051)***
contexto_poblacion			-0.173 (0.025)***
contexto_robos_poblacion			-0.020 (0.017)
contexto_pobreza_personas			0.058 (0.017)***
contexto_modelo_p_robos			-0.300 (0.019)***
registro_seccional_region_cuyo			-0.423 (0.092)***
registro_seccional_region_GBA			-0.333 (0.066)***
registro_seccional_region_noroeste			0.616 (0.078)***
registro_seccional_region_pampeana			-0.491 (0.060)***
registro_seccional_region_patagonia			-0.306 (0.079)***
Num.Obs.	31866	31866	31866
AIC	36821.4	36226.3	35443.1
BIC	36896.7	36335.1	35627.2
Log.Lik.	-18401.702	-18100.142	-17699.529

Estimación en base a datos de bajas DNRPA

Modelo base (baseline) de regresión logística seleccionado:

$$\log\left[\frac{P(..\text{y} = \widehat{\text{corta_duracion}})}{1 - P(..\text{y} = \widehat{\text{corta_duracion}})}\right] = 1.25 +$$
$$0.26(\text{titular_edad_inscripcion}) -$$
$$0.07(\text{titular_titularidad_compartida}) -$$
$$0.17(\text{contexto_poblacion}) -$$
$$0.02(\text{contexto_robos_poblacion}) +$$
$$0.06(\text{contexto_pobreza_personas}) -$$
$$0.3(\text{contexto_modelo_p_robos}) +$$
$$0.42(\text{automotor_PC1}) -$$
$$0.4(\text{automotor_PC2}) -$$
$$0.7(\text{automotor_PC3}) -$$
$$0.18(\text{robos_provincia}) -$$
$$1.39(\text{automotor_categoria_AUTOMOVIL}) -$$
$$1.88(\text{automotor_categoria_CAMIONETA}) +$$
$$0.32(\text{titular_genero_Femenino}) +$$
$$0.39(\text{titular_tipo_persona_Jurídica}) -$$
$$0.42(\text{registro_seccional_region_cuyo}) -$$
$$0.33(\text{registro_seccional_region_GBA}) +$$
$$0.62(\text{registro_seccional_region_noroeste}) -$$
$$0.49(\text{registro_seccional_region_pampeana}) -$$
$$0.31(\text{registro_seccional_region_patagonia}) -$$
$$1(\text{automotor_cluster_X1}) -$$
$$2.04(\text{automotor_cluster_X2})$$

Reporte de la mentoría

Trabajo Final de Especialización de Karina Bartolomé:

"Factores determinantes de la tenencia de automóviles en Argentina.
Un análisis multivariado de los registros del automotor para la gestión eficiente de datos en la Asociación de Fábricas Automotores."

Mentora: Natalia Salaberry
Fecha: diciembre 2021

La alumna Karina Bartolomé en su trabajo final de especialización logra desarrollar de manera concisa y acabada su proyecto inicial de trabajo. Plantea de manera clara la existencia de una necesidad en una organización que luego resuelve con las herramientas aprendidas en el campo diciplinar de la especialización en Métodos Cuantitativos para la Gestión y Análisis de Datos en Organizaciones. En este sentido, el problema propuesto a resolver se vincula a la detección de factores que caracterizan la demanda de automóviles en Argentina, contribuyendo a que la organización Asociación de Fábricas de Automotores (ADEFA) pueda obtener conocimiento acerca de las preferencias de los titulares para avanzar en aumentar la competitividad de las empresas que nuclea.

Para cumplimentar con el objetivo planteado, en un primer apartado desarrolla una clara contextualización del sector en Argentina al cual pertenece la organización. A partir de este punto, describe la situacionalidad en base a las oportunidades y desafíos que presenta la disponibilidad de grandes volúmenes de datos. Pero también advierte sobre la necesidad e importancia de una gestión adecuada de estos para lograr la construcción de valor agregado para la toma de decisiones.

En un segundo apartado, se centra en exponer de manera concisa las herramientas metodológicas a utilizar y que fueron aprendidas en diversas materias de la especialización siendo lo que le permite identificar los factores relevantes en la demanda de automóviles cumplimentando con su objetivo principal. Desde el aspecto técnico de implementación de modelos predictivos, realiza un trabajo perfecto, respetando todas las etapas de aplicación, análisis y evaluación de resultados. Supo seleccionar adecuadamente los modelos acordes que le permitieron cumplimentar con el objetivo buscado.

Finalmente, en un tercer apartado expone como los resultados obtenidos permiten generar valor agregado para la toma de decisiones a partir de desarrollar un análisis de manera

completa. Al mismo tiempo, advierte sobre la necesidad de una gestión de datos más eficiente que permitirá propiciar mayor volumen de información para ser procesada y analizada a fin de obtener resultados aún más favorables. Expone claramente como podría obtenerse datos relevantes para determinar acabadamente otros factores que contribuyen a profundizar en el conocimiento de la demanda de automóviles.

De esta manera, la alumna logra cumplimentar con el objetivo principal y los específicos de forma ordenada y clara, realizando un trabajo completamente estructurado en el marco de las pautas establecidas. Existe una coherencia entre la problemática planteada, el título y las palabras claves lo que le permitió poder elaborar un planteo y desarrollo adecuado.

Un aspecto para mejorar en futuros trabajos que se le sugiere es en cuanto a la forma de escritura, profundizar más en la construcción de párrafos analíticos y concluyentes de forma tal que exista un relato continúo.