

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS MATEMÁTICAS
Grado en Matemáticas y Ciencia de Datos

**Aplicación de técnicas de *Machine Learning*
dentro del ámbito de la Fórmula 1**



Trabajo de Fin de Grado
Curso 2025/2026

Autor: **Álvaro Pérez Ramos**

Tutor: **Daniel Vélez Serrano**

Resumen

Este Trabajo de Fin de Grado predice el tiempo por vuelta de un piloto durante el segundo *stint* de un Gran Premio de Fórmula 1, emitiendo la predicción en el instante de la primera parada en boxes y empleando exclusivamente la información disponible en ese momento. A partir de datos públicos de cronometrage (OpenF1, temporadas 2023 a 2025) se construye una tabla de 17 093 vueltas agrupadas en 771 unidades piloto–carrera, con cincuenta y cuatro variables explicativas. Se comparan tres definiciones alternativas de la variable de respuesta —el nivel del tiempo por vuelta, ese mismo nivel *anclado* al ritmo mediano del primer *stint*, y su incremento a dos vueltas vista— y dos familias de modelos, un modelo lineal generalizado y un *gradient boosting*. La aportación metodológica central es mostrar, teórica y empíricamente, que la definición del objetivo condiciona qué familia de modelos puede resolverlo: se demuestra que el anclaje deja invariantes las predicciones de un modelo lineal que incluya el ancla entre sus regresores, y se comprueba que reduce un 39,5 por ciento el error de un modelo de árboles, que no extrapola. Los modelos se ajustan con las temporadas 2023 y 2024 y se evalúan sobre la de 2025, de modo que la evaluación es estrictamente posterior al ajuste.

Abstract

This Bachelor’s Thesis predicts a driver’s lap time during the second *stint* of a Formula 1 Grand Prix, issuing the prediction at the moment of the first pit stop and using only the information available at that instant. From public timing data (OpenF1, 2023 to 2025 seasons), a modelling table is built with 17 093 laps grouped into 771 driver–race units and fifty-four explanatory variables. Three alternative definitions of the response variable are compared —the lap time level, that same level *anchored* to the median pace of the first *stint*, and its increment two laps ahead— together with two families of models, a generalised linear model and a *gradient boosting*. The central methodological contribution is to show, theoretically and empirically, that the definition of the target determines which family of models can solve it: anchoring is proved to leave unchanged the predictions of a linear model that includes the anchor among its regressors, and is shown to reduce by 39.5 per cent the error of a tree-based model, which cannot extrapolate. The models are fitted on the 2023 and 2024 seasons and evaluated on the 2025 season, so that evaluation is strictly subsequent to fitting.

Índice general

1. Introducción	1
1.1. Motivación	1
1.2. Trabajos relacionados	1
1.3. Objetivo del trabajo	2
1.4. Estructura del documento	2
2. El problema	4
2.1. La carrera y la vuelta	4
2.2. El neumático, el stint y la parada	4
2.2.1. Compuestos	4
2.2.2. Stint	5
2.2.3. La parada en boxes	5
2.3. Qué hace variar el tiempo por vuelta	5
2.3.1. La degradación	5
2.3.2. El combustible	6
2.3.3. El aire sucio	6
2.4. La estrategia	6
2.4.1. El problema de decisión	6
2.4.2. <i>Undercut</i> y <i>overcut</i>	7
2.4.3. Lo que el modelo no puede ver	7
2.5. Interrupciones y vueltas no representativas	7
2.6. Formulación del problema	8
3. Preliminares y metodología	9
3.1. Preliminares: modelos utilizados	9
3.1.1. Modelos lineales generales	9
3.1.2. Modelos de tipo <i>boosting</i>	13
3.1.3. Predictores de referencia	17
3.2. Metodología	18
3.2.1. Planteamiento	18
3.2.2. Metodologías de modelización predictiva	18
3.2.3. Generalización y partición de la muestra	19
3.2.4. Construcción de la tabla de entrada e imposición de filtros	20
3.2.5. Definición de la respuesta y de las variables explicativas	20
3.2.6. Análisis exploratorio	21
3.2.7. Tratamiento de variables	22
3.2.8. Ajuste de los modelos	22
3.2.9. Validación e interpretación de resultados	23

4. Aplicación de la metodología	24
4.1. Fuente de datos	24
4.2. Software utilizado	24
4.3. Aplicación de la metodología	25
4.3.1. Construcción de la tabla de entrada e imposición de filtros	25
4.3.2. Definición de la variable a predecir y generación de variables explicativas	26
4.3.3. Análisis exploratorio	28
4.3.4. Tratamiento de variables	30
4.3.5. Ajuste de modelos analíticos	32
4.3.6. Validación de resultados	33
4.3.7. Interpretación de los resultados del modelo ganador	35
4.4. Limitaciones	38
5. Conclusiones	39
5.1. Conclusiones	39
5.2. Líneas futuras	40
A. Comparación de tratamientos de valores atípicos	41
B. Optimización de hiperparámetros	43
C. Detalle de los valores atípicos	44
D. Valores exactos de la interpretación	45
E. Dependencia de los valores de Shapley	47
F. Definición de las variables explicativas	48
G. Descriptivo univariante completo	51
H. Descriptivo bivariante completo	53
I. Capacidad predictiva sobre los tres objetivos	54

Capítulo 1

Introducción

1.1. Motivación

Una carrera de Fórmula 1 se decide, en buena medida, fuera de la pista. El reglamento obliga a montar al menos dos compuestos distintos de neumático, de modo que toda carrera en seco incluye al menos una parada en boxes que cuesta entre veinte y treinta segundos. Elegir cuándo parar y qué compuesto montar es una decisión que se toma una sola vez, sin posibilidad de rectificar, y cuyo acierto depende de anticipar cómo se comportará el coche durante las decenas de vueltas siguientes.

Esa anticipación es un problema de predicción bien planteado. En el instante de la parada se dispone de información abundante y reciente, un tramo completo de vueltas cronometradas del propio coche, ese día y en ese circuito, y de un horizonte concreto sobre el que decidir. La disponibilidad pública de datos de cronometraje vuelta a vuelta permite además abordarlo con métodos estadísticos sin acceso a información privada de los equipos.

1.2. Trabajos relacionados

La literatura sobre la toma de decisiones estratégicas en competición automovilística puede estructurarse en tres grandes líneas de investigación, que abarcan desde la simulación de estrategias de carrera hasta la predicción de las variables que las sustentan y su optimización mediante técnicas de aprendizaje automático.

La primera emplea **simulación de eventos discretos** para evaluar planes de carrera completos. Bekker y Lotz [4] construyen un simulador de carrera de Fórmula 1 con el que comparan estrategias de parada bajo distintos escenarios, y muestran que el número óptimo de paradas depende críticamente de la pérdida de boxes del circuito y del ritmo de degradación. Heilmeyer y colaboradores extienden esa idea en dos trabajos: primero con un simulador determinista orientado a la decisión estratégica [18], y después incorporando el carácter aleatorio de los coches de seguridad, los fallos mecánicos y la dispersión de los tiempos por vuelta mediante métodos de Monte Carlo [19]. La conclusión común de esta línea es que la estrategia óptima es muy sensible a la incertidumbre, de modo que una recomendación basada únicamente en valores esperados resulta frágil.

La segunda línea aborda directamente la **predicción de las variables** utilizadas en esas simulaciones. Tulabandhula y Rudin [37] estudian, sobre datos de NASCAR, la predicción del efecto del cambio de neumáticos y del rendimiento bajo bandera amarilla, y documentan una dificultad que reaparece en este trabajo: las decisiones de los equipos condicionan qué observaciones se llegan a registrar, lo que sesga cualquier estimación ingenua. Más recientemente, Todd y colaboradores [36] predicen la energía acumulada por el neumático en Fórmula 1 mediante modelos de series temporales explicables, y Sasikumar y colaboradores [31] proponen

un sistema de apoyo a la decisión de parada basado en aprendizaje profundo.

La tercera línea combina **aprendizaje automático con búsqueda** para optimizar la estrategia. Liu y Fotouhi [23] emplean redes neuronales junto a búsqueda en árbol de Monte Carlo en Fórmula E, y Choo [12] desarrolla modelos de decisión en tiempo real para carreras profesionales. Anandatama y Gamayanto [2] aplican *gradient boosting* con optimización bayesiana de hiperparámetros a la predicción del rendimiento del neumático en un Gran Premio concreto, en una línea metodológica próxima a la de este trabajo.

De este repaso se desprenden dos limitaciones que motivan el enfoque adoptado aquí. El primero es que la mayor parte de los trabajos citados predicen el tiempo por vuelta como una variable auxiliar de un simulador, sin analizar con detalle cómo debe definirse la variable objetivo ni qué familia de modelos conviene según esa definición. El segundo es que rara vez se comparan los modelos complejos contra referencias verdaderamente sencillas, de modo que no queda claro cuánto de la capacidad predictiva procede del modelo y cuánto de la inercia del propio fenómeno.

1.3. Objetivo del trabajo

El objetivo de este trabajo es **predecir el tiempo por vuelta de un piloto durante el segundo *stint* de un Gran Premio de Fórmula 1**, realizando la predicción en el instante en que el coche completa su primera parada en boxes y empleando exclusivamente la información disponible en ese momento.

Ese objetivo general se concreta en cuatro objetivos específicos:

- Construir, a partir de datos públicos de cronometraje, una tabla de modelado con una definición precisa de la unidad de observación y de las variables explicativas disponibles en el instante de la predicción.
- Comparar dos definiciones alternativas de la variable objetivo, el nivel del tiempo por vuelta y su incremento a dos vueltas vista, y determinar cuál resulta más adecuada para cada familia de modelos.
- Contrastar la capacidad predictiva de un modelo lineal generalizado frente a un modelo de tipo *boosting*, ambos evaluados sobre una temporada posterior a la empleada en el ajuste, y situarlos frente a predictores de referencia deliberadamente sencillos.
- Interpretar el modelo ganador, contrastando lo que ha aprendido con las relaciones que el análisis exploratorio permite anticipar.

Queda fuera del alcance del trabajo la optimización de la estrategia completa de carrera, que exigiría modelar la posición en pista, los adelantamientos y la aparición de coches de seguridad.

1.4. Estructura del documento

Tras la introducción realizada en este primer capítulo, el capítulo 2 describe el problema y la terminología de la Fórmula 1 necesaria para seguir el resto del documento: qué es un *stint*, qué gobierna la evolución del tiempo por vuelta y en qué consisten las maniobras estratégicas asociadas a la parada.

El capítulo 3 presenta los preliminares y la metodología. La sección 3.1 describe las familias de modelos empleadas, modelos lineales generalizados y modelos de tipo *boosting*, con su formulación, sus hiperparámetros y la forma en que se interpretan. La sección 3.2 expone la

metodología de modelización predictiva seguida, en seis fases y sin referencia al problema de aplicación.

El capítulo 4 aplica esa metodología al problema concreto: describe la fuente de datos y el software empleado y recorre las seis fases, desde la construcción de la tabla de entrada hasta la interpretación del modelo ganador mediante valores de Shapley.

Finalmente, el capítulo 5 resume el trabajo realizado, recoge las conclusiones alcanzadas y propone líneas de continuación. Se incluyen además dos anexos con los resultados detallados del tratamiento de valores atípicos y de la optimización de hiperparámetros.

Capítulo 2

El problema

Este capítulo describe el problema real y fija el vocabulario que se empleará en los dos capítulos siguientes. Se ha situado antes del desarrollo matemático por una razón práctica: buena parte de las decisiones de modelado del capítulo 4 solo se entienden como respuesta a hechos concretos del dominio, y conviene que esos hechos estén enunciados antes de que se invoquen. La tabla 2.1 recoge, al final, los términos técnicos con su definición.

2.1. La carrera y la vuelta

Un Gran Premio de Fórmula 1 es una carrera de entre 44 y 78 vueltas a un circuito cerrado, disputada por veinte coches de diez escuderías, a dos coches por escudería. La unidad de medida fundamental es el *tiempo por vuelta*: la duración, en segundos, que un coche emplea en completar una vuelta completa al trazado. Es la variable que este trabajo predice.

El tiempo por vuelta oscila entre unos 70 segundos en los circuitos más cortos y unos 115 en los más largos, de modo que la variación entre circuitos es de más de cuarenta segundos. Dentro de una misma carrera, en cambio, la variación relevante es de décimas: dos coches separados por medio segundo por vuelta están, en términos deportivos, en ligas distintas. Esta desproporción entre la variación entre carreras y la variación dentro de una carrera es el hecho que gobierna la definición del objetivo en la sección 4.3.2.

El fin de semana de carrera consta de varias sesiones. Los *entrenamientos libres* permiten a los equipos rodar y probar reglajes, sin consecuencia deportiva directa. La *clasificación* determina el orden de salida, o *parrilla*, según el mejor tiempo por vuelta de cada piloto. La *carrera* es la sesión puntuable. En algunos fines de semana se disputa además una *carrera al esprint*, de menor distancia, lo que reduce el número de sesiones de entrenamientos libres de tres a una.

2.2. El neumático, el stint y la parada

2.2.1. Compuestos

El neumático es el único elemento del coche que toca el asfalto y, por tanto, el que determina el agarre disponible. Pirelli, proveedor único, fabrica una gama de compuestos de seco numerados de *C1* a *C6*, del más duro al más blando. Un compuesto blando ofrece más agarre y, por tanto, menor tiempo por vuelta, pero se degrada antes; uno duro, al contrario.

A cada Gran Premio, Pirelli lleva únicamente tres compuestos de esa gama, que etiqueta localmente como *blando*, *medio* y *duro*. La consecuencia, fácil de pasar por alto, es que esas etiquetas son *relativas*: el medio de un circuito no es el mismo caucho que el medio de otro.

El medio de Monza puede ser un C4 y el de Suzuka un C2. Por eso el trabajo registra tanto la etiqueta local, en escala ordinal de uno a tres, como el número absoluto de la gama.

El reglamento obliga a emplear al menos dos compuestos distintos a lo largo de la carrera en seco. De ahí se sigue un hecho que hace bien planteado todo el problema: **toda carrera en seco tiene, como mínimo, una parada en boxes obligatoria.**

2.2.2. Stint

Se llama *stint* a un tramo de vueltas consecutivas completadas con el mismo juego de neumáticos. Una carrera se descompone, por tanto, en tantos *stints* como paradas realice el piloto más uno: quien para una sola vez completa dos *stints*, quien para dos veces completa tres.

Este trabajo se ocupa del *segundo stint*: el tramo que empieza cuando el coche sale de boxes tras su primera parada. La razón de esa elección es la información disponible en ese instante, y se detalla más abajo.

2.2.3. La parada en boxes

Para cambiar los neumáticos, el coche abandona el trazado, recorre el *pit lane* —el carril de boxes— a velocidad limitada por reglamento, normalmente 80 kilómetros por hora, se detiene unos dos segundos mientras los mecánicos cambian las cuatro ruedas, y reincorpora al trazado.

El coste de esa operación, medido como diferencia entre el tiempo de la vuelta en que se entra a boxes y el que ese mismo coche habría hecho rodando normalmente, se denomina *pérdida de boxes* y se denotará Δ_{box} . Su magnitud está entre veinte y treinta segundos y varía apreciablemente entre circuitos, según la longitud y la geometría del acceso. Esta variación es una de las razones por las que la estrategia óptima no es la misma en todos los trazados.

Se llaman *vuelta de entrada* a la que termina con el coche entrando en boxes y *vuelta de salida* a la siguiente. Ninguna de las dos informa sobre el ritmo del coche, puesto que ambas incluyen el recorrido a velocidad limitada; por eso el capítulo 4 las descarta.

2.3. Qué hace variar el tiempo por vuelta

Dentro de un *stint*, y en ausencia de incidentes, el tiempo por vuelta de un coche está gobernado por dos fuerzas de signo contrario y magnitud comparable.

2.3.1. La degradación

El neumático pierde prestaciones con el uso. La goma se desgasta, la temperatura de trabajo se desplaza fuera de su ventana óptima y el agarre disminuye. Se llama *degradación* a esa pérdida y se mide como el incremento de tiempo por vuelta por cada vuelta de uso del juego, en segundos por vuelta.

La degradación no es constante a lo largo de un *stint*. Suele distinguirse una fase inicial de estabilización, en la que el neumático alcanza su temperatura de trabajo; una fase central de degradación aproximadamente lineal; y, en algunos casos, lo que en la jerga se conoce como el *precipicio*: una caída brusca del rendimiento cuando la banda de rodadura se agota. El precipicio hace que prolongar un *stint* más allá de la vida útil efectiva del neumático pueda provocar una pérdida considerable de rendimiento y, por ello, condiciona de forma importante las decisiones sobre la duración del *stint*.

Se llama *edad del neumático* al número de vueltas completadas con el juego montado. Es la variable frente a la que se mide habitualmente la degradación, por ser la que cuenta el uso acumulado del juego.

2.3.2. El combustible

El coche parte con el depósito lleno —hasta 110 kilogramos de combustible— y lo consume a ritmo prácticamente constante. Como el tiempo por vuelta crece con la masa del coche, el efecto es que el coche va ganando tiempo a lo largo de la carrera por el mero hecho de aligerarse: del orden de tres a cuatro centésimas por vuelta.

Este efecto va en dirección contraria a la degradación y es de magnitud comparable. La consecuencia práctica es que el tiempo por vuelta observado dentro de un *stint* puede resultar plano, creciente o decreciente según cuál de los dos domine. La consecuencia estadística, que se desarrolla en la sección 4.4, es más incómoda: dentro de un *stint* la edad del neumático y la vuelta de carrera avanzan a la vez, de modo que sus efectos no son separables.

2.3.3. El aire sucio

Un Fórmula 1 genera la mayor parte de su agarre por efecto aerodinámico: el flujo de aire sobre alerones y fondo plano presiona el coche contra el asfalto. Al rodar detrás de otro coche se circula dentro de su estela —aire turbulento y ya acelerado, con menor presión dinámica—, lo que se conoce como *aire sucio*. El coche perseguidor pierde carga aerodinámica y, con ella, velocidad de paso por curva. A esa pérdida se suman dos efectos secundarios en la misma dirección: el radiador recibe aire más caliente, lo que obliga a levantar el pie del acelerador para proteger las temperaturas, y el neumático desliza más, lo que acelera su desgaste.

Tres características del fenómeno resultan especialmente relevantes para el modelado. En primer lugar, su efecto depende de la distancia, ya que la estela se disipa a medida que aumenta el espacio respecto al coche precedente. Por tanto, no debe entenderse como un efecto binario, sino como una función continua del hueco entre ambos coches. En segundo lugar, su magnitud depende del circuito: los trazados con curvas rápidas y una elevada exigencia aerodinámica pueden amplificar la pérdida de rendimiento, mientras que en las rectas el efecto puede ser incluso favorable. En estas últimas, el *rebufo* —la reducción de la resistencia al avance que experimenta un coche al circular en la zona de baja presión generada por el vehículo precedente— permite alcanzar una mayor velocidad punta. Así, un mismo fenómeno puede perjudicar el rendimiento en curva y favorecerlo en recta. El *DRS*, mediante la apertura temporal de un elemento móvil del alerón trasero en las zonas habilitadas por el reglamento, puede reducir parcialmente esta pérdida de rendimiento en recta. Por último, el efecto puede propagarse a lo largo de varios coches. Cuando varios vehículos quedan agrupados detrás de uno que marca un ritmo inferior, se forma lo que se conoce como un *tren*. En este contexto, ocupar la tercera posición del tren no equivale a ocupar la segunda, ya que, además del efecto de la estela, el ritmo de los coches que lo integran queda condicionado por el vehículo que circula delante.

2.4. La estrategia

2.4.1. El problema de decisión

En el instante de la primera parada, el muro de boxes ha observado ya un *stint* completo: conoce el ritmo real de su coche ese día, en esa pista, con ese piloto y con ese nivel de combustible. Y debe decidir dos cosas simultáneamente: qué compuesto montar y cuántas vueltas estirarlo. Todo lo que ocurra en el resto de la carrera depende de esa elección, tomada además con el coche detenido y sin margen para reconsiderarla.

Esta asimetría —mucha información acumulada y una única oportunidad de usarla— es la que convierte el instante de la parada en un problema de predicción bien planteado, y la razón de que sea el momento desde el que se emiten todas las predicciones de este trabajo.

2.4.2. *Undercut y overcut*

Dos maniobras estratégicas merecen definición porque ilustran por qué predecir el tiempo por vuelta del segundo *stint* tiene valor operativo.

El *undercut* consiste en parar antes que un rival al que se persigue. Al salir con neumático fresco mientras el otro sigue con el gastado, se rueda más rápido durante las vueltas que el rival tarda en parar, y si esa ganancia acumulada supera la diferencia previa, se le adelanta sin necesidad de disputarle la posición en pista.

Formalmente: sean A el perseguidor y B el perseguido, separados en pista por una diferencia de tiempo de g segundos. Si A para en la vuelta L y B en la vuelta $L + k$, ambos pagan, en segundos, la misma pérdida de boxes Δ_{box} , que por tanto se cancela. Durante esas k vueltas, A rueda con neumático nuevo a un ritmo $p_{\text{nuevo}}(j)$ y B con el gastado a $p_{\text{gastado}}(j)$. El *undercut* tiene éxito si

$$\sum_{j=1}^k \left[p_{\text{gastado}}(j) - p_{\text{nuevo}}(j) \right] > g. \quad (2.1)$$

La expresión (2.1) deja ver por qué el problema de este trabajo es el que es: lo que decide la maniobra no son los tiempos absolutos, sino la *diferencia* de ritmo entre un neumático nuevo y uno gastado, acumulada a lo largo de unas pocas vueltas. Predecir esa diferencia es exactamente el objetivo T_2 de la sección 4.3.2, y predecir la curva completa de p_{nuevo} es el objetivo T_1 .

El *overcut* es la maniobra simétrica: retrasar la parada respecto al rival. Funciona cuando el neumático propio aguanta bien y el rival, tras parar, sale a tráfico y no puede aprovechar su goma nueva.

2.4.3. Lo que el modelo no puede ver

Existe un conjunto de razones por las que un equipo detiene un coche que no queda registrado en ningún dato público: una presión de neumático fuera de ventana, un aviso de ampollamiento, un pinchazo lento, daños en el alerón delantero, o sencillamente la necesidad de cubrir la parada de un rival directo. Conviene tenerlo presente: cuando este trabajo compare lo predicho con lo ocurrido, el equipo decidió con más información que el modelo, no con menos.

2.5. Interrupciones y vueltas no representativas

No todas las vueltas de una carrera informan sobre el ritmo del coche. El reglamento prevé varios regímenes de neutralización, señalizados con banderas.

La *bandera verde* indica carrera en condiciones normales, y es el único régimen en el que el tiempo por vuelta refleja el ritmo del coche. La *bandera amarilla* obliga a levantar el pie del acelerador en un sector concreto. El *coche de seguridad* agrupa a todo el pelotón detrás de un vehículo a velocidad reducida, con lo que los tiempos por vuelta se disparan y las diferencias entre coches se anulan; su variante virtual, el *VSC*, impone un tiempo mínimo por sector sin agrupar físicamente. La *bandera roja* detiene la carrera; tras ella se reanuda con salida parada o lanzada, lo que reinicia la numeración de *stints*.

A ello se añaden las vueltas iniciales, con el pelotón aún comprimido tras la salida, y las de entrada y salida de boxes ya descritas.

Todas estas vueltas contienen variación de tiempo que no procede del coche sino de la situación de carrera, y que además está correlacionada con la posición. Incluirlas contaminaría cualquier estimación de degradación y sesgaría también los efectos de tráfico. De ahí los filtros que se imponen a la hora de modelar y que son detallados en la sección 4.3.1.

2.6. Formulación del problema

Reuniendo lo anterior, el problema que aborda este trabajo se enuncia así.

En el instante en que un piloto completa su primera parada en boxes y reincorpora a pista, se dispone de: todas las vueltas de su primer *stint*, con lo que se conoce su ritmo real ese día; el compuesto que se acaba de montar; el estado del pelotón a su alrededor; las condiciones ambientales; y los resultados de clasificación y entrenamientos libres. Se desea predecir cómo evolucionará su tiempo por vuelta a lo largo del segundo *stint*, sin observar ninguna vuelta intermedia.

Se abordan dos formulaciones de esa pregunta, que responden a necesidades operativas distintas y que el capítulo 4 define con precisión: predecir el *nivel* del tiempo por vuelta en cada vuelta del segundo *stint*, y predecir su *incremento* a dos vueltas vista, que es la magnitud que interviene en (2.1).

CUADRO 2.1: Glosario de términos del dominio empleados en los capítulos siguientes.

Término	Definición
<i>Stint</i>	Tramo de vueltas consecutivas completadas con el mismo juego de neumáticos.
Compuesto	Tipo de goma del neumático. Gama de seco C1 (más duro) a C6 (más blando), etiquetada localmente en cada Gran Premio como duro, medio y blando.
Degradación	Pérdida de prestaciones del neumático con el uso, medida en segundos por vuelta y por vuelta de edad del juego.
Precipicio	Caída brusca del rendimiento del neumático al agotarse la banda de rodadura.
Edad del neumático	Número de vueltas completadas con el juego montado.
Pérdida de boxes	Tiempo perdido en una parada respecto a rodar en pista, entre veinte y treinta segundos según el circuito.
<i>Pit lane</i>	Carril de acceso a boxes, recorrido a velocidad limitada.
Vuelta de entrada / salida	Vuelta que termina con el coche entrando en boxes / vuelta siguiente a la parada.
Aire sucio	Estela turbulenta de un coche precedente, que reduce la carga aerodinámica del perseguidor.
Rebufo	Reducción de resistencia al avance por circular tras otro coche; a diferencia del aire sucio, beneficia en recta.
DRS	Alerón trasero móvil, autorizado a menos de un segundo del precedente en zonas determinadas, que reduce la resistencia en recta.
Tren	Cadena de coches consecutivos separados por menos de un segundo y medio, cuyo ritmo lo impone el más lento de la cabeza.
<i>Undercut</i>	Adelantar a un rival parando antes que él y aprovechando la ventaja del neumático nuevo, según (2.1).
<i>Overcut</i>	Maniobra simétrica: retrasar la parada respecto al rival.
Bandera verde	Carrera en condiciones normales; único régimen en el que el tiempo por vuelta refleja el ritmo del coche.
Coche de seguridad	Neutralización que agrupa al pelotón a velocidad reducida. Su variante virtual, el VSC, impone un tiempo mínimo por sector.
Bandera roja	Interrupción de la carrera; al reanudarse reinicia la numeración de <i>stints</i> .
Parrilla	Orden de salida, determinado por la clasificación.
Carrera al esprint	Carrera corta adicional en algunos fines de semana; reduce las sesiones de entrenamientos libres de tres a una.
Pelotón	Conjunto de los veinte coches en pista, y por extensión la configuración de huecos entre ellos en un instante dado.

Capítulo 3

Preliminares y metodología

3.1. Preliminares: modelos utilizados

Se dispone de una muestra $\{(x_i, y_i)\}_{i=1}^n$, donde $y_i \in \mathbb{R}$ es la variable de respuesta y $x_i \in \mathbb{R}^p$ el vector de variables explicativas o regresores de la observación i . Se denota por $X \in \mathbb{R}^{n \times (p+1)}$ la matriz de regresores, cuya primera columna es de unos, y por $y \in \mathbb{R}^n$ el vector de respuestas.

Ajustar o *entrenar* un modelo consiste en elegir, dentro de una clase de funciones $\mathcal{F}$, aquella que minimiza el error cometido sobre la muestra:

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i)), \quad (3.1)$$

donde $L : \mathbb{R}^2 \rightarrow [0, \infty)$ es la función de pérdida, que mide el coste de predecir $f(x_i)$ cuando el valor observado es y_i . Las dos decisiones que definen un método son, por tanto, la clase $\mathcal{F}$ y la pérdida L .

Conviene recordar el resultado que ordena toda la comparación posterior. Si $L(y, f) = (y - f)^2$, el minimizador poblacional de (3.1) es la esperanza condicional $f^*(x) = \mathbb{E}[y \mid x]$; si $L(y, f) = |y - f|$, es la mediana condicional. Ambos difieren cuando la distribución condicional es asimétrica, de modo que la pérdida determina qué característica de esa distribución se estima y no solo con qué rapidez se converge a ella.

3.1.1. Modelos lineales generales

3.1.1.1. Formulación y elementos

El modelo lineal general postula

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i, \quad i = 1, \dots, n, \quad \text{o bien} \quad y = X\beta + \varepsilon, \quad (3.2)$$

con las hipótesis

$$\mathbb{E}[\varepsilon \mid X] = 0, \quad \mathbb{V}[\varepsilon \mid X] = \sigma^2 I_n, \quad \text{rango}(X) = p + 1, \quad (3.3)$$

donde I_n denota la matriz identidad de orden n . Los elementos de (3.2) son: el vector de respuesta $y \in \mathbb{R}^n$; la matriz de regresores X , cuya columna j recoge los valores de la variable explicativa j ; el vector de parámetros $\beta \in \mathbb{R}^{p+1}$, desconocido y objeto de la estimación; y el vector de perturbaciones $\varepsilon \in \mathbb{R}^n$.

La primera hipótesis de (3.3) garantiza insesgadez. La segunda impone que las perturbaciones tengan varianza constante (homocedasticidad) y estén incorreladas entre sí. La tercera,

que X tenga rango completo, es condición necesaria y suficiente para que β quede identificado de forma única.

El calificativo *lineal* se refiere a la aditividad de los efectos de los regresores en el modelo: cada variable contribuye a la respuesta con una cantidad proporcional a su valor y esas contribuciones se suman. Nada impide, en cambio, que las columnas de X contengan transformaciones no lineales de las variables originales, como x^2 , $\log x$ o el producto x_1x_2 .

3.1.1.2. Estimación por mínimos cuadrados

El estimador de mínimos cuadrados ordinarios resuelve $\hat{\beta} = \arg \min_b \|y - Xb\|_2^2$, donde b recorre los posibles valores del vector de parámetros. Calculando el gradiente de esa suma de cuadrados e igualándolo a cero se obtienen las ecuaciones normales $X'Xb = X'y$ y, bajo rango completo,

$$\hat{\beta} = (X'X)^{-1}X'y, \quad \hat{y} = X\hat{\beta} = Hy, \quad H = X(X'X)^{-1}X', \quad (3.4)$$

donde H es la matriz de proyección ortogonal sobre el subespacio generado por las columnas de X . De (3.3) y (3.4) se sigue $\mathbb{E}[\hat{\beta}] = \beta$ y $\mathbb{V}[\hat{\beta}] = \sigma^2(X'X)^{-1}$. El teorema de Gauss–Markov establece que $\hat{\beta}$ tiene varianza mínima dentro de la clase de los estimadores lineales insesgados [33]. Bajo normalidad de ε , coincide además con el estimador de máxima verosimilitud.

3.1.1.3. Invariancia frente al anclaje del objetivo

Una consecuencia menos evidente de (3.4) conviene enunciarse aquí, porque delimita para qué familias de modelos supone una diferencia restar a la respuesta una cantidad conocida de antemano —operación que la sección 3.2.5 denominará *anclaje*— y para cuáles ninguna.

Proposición 3.1 (Invariancia del modelo lineal frente al anclaje). *Supóngase que el ancla figura entre las variables explicativas, es decir, que el vector $(\hat{a}_1, \dots, \hat{a}_n)'$ coincide con la columna k -ésima de la matriz de regresores. Sean $\hat{y}$ las predicciones de la regresión por mínimos cuadrados de y sobre X , y $\hat{\tilde{y}}$ las de la regresión de $y - Xe_k$ sobre X , siendo e_k el k -ésimo vector de la base canónica. Entonces $\hat{\tilde{y}} + Xe_k = \hat{y}$, y los residuos de ambas regresiones coinciden.*

Demostración. Por (3.4), $\hat{\tilde{y}} = H(y - Xe_k) = Hy - HXe_k$. Como H proyecta sobre el espacio de columnas de X y Xe_k pertenece a él, $HXe_k = Xe_k$, de donde $\hat{\tilde{y}} = \hat{y} - Xe_k$. La igualdad de residuos es inmediata: $(y - Xe_k) - \hat{\tilde{y}} = y - \hat{y}$. $\square$

Para un modelo lineal que incluya el ancla entre sus regresores, anclar es por tanto una operación inocua: el estimador la absorbe desplazando en una unidad el coeficiente correspondiente, y el error de predicción no cambia. Para un modelo de árboles la situación será la contraria, por un motivo que la proposición 3.3 precisará más adelante y que conviene anticipar para evitar una lectura equivocada. No se tratará de que los árboles no dispongan del ancla: la tendrán entre sus regresores, exactamente igual que el modelo lineal. Se tratará de que solo pueden emplearla para *partir* el espacio, nunca para *sumarla* a la predicción. Un árbol puede decidir que, si el ancla supera cierto umbral, la predicción sea una constante determinada; pero no puede representar la función «ancla más una corrección», que es precisamente lo que un modelo lineal estima cuando asigna a esa columna un coeficiente próximo a uno.

3.1.1.4. Interpretación mediante los parámetros

El coeficiente β_j es la derivada parcial $\partial \mathbb{E}[y \mid x] / \partial x_j$: mide el cambio esperado en la respuesta ante un incremento unitario de x_j manteniendo constantes los restantes regresores. Esta es la propiedad que convierte a los modelos lineales en la familia interpretable por excelencia, y la razón de conservarlos en la comparación aunque no resulten los más precisos.

La formalización de la cláusula «manteniendo constantes los restantes regresores» es el teorema de Frisch–Waugh–Lovell [16, 24].

Teorema 3.2 (Frisch–Waugh–Lovell). *Sea $X = [X_1 \ X_2]$ una partición por bloques de la matriz de regresores y sea $M_2 = I_n - X_2(X_2'X_2)^{-1}X_2'$. El subvector $\hat{\beta}_1$ de (3.4) coincide con el estimador de mínimos cuadrados de la regresión de M_2y sobre M_2X_1 .*

Dicho sin matrices, el coeficiente $\hat{\beta}_j$ mide el efecto de x_j sobre la respuesta una vez descontado el efecto que pueden explicar las demás variables del modelo. Para entender la precisión con la que se estima este coeficiente, resulta útil imaginar que primero se elimina de x_j la parte que puede explicarse mediante los demás regresores. La estimación de $\hat{\beta}_j$ se apoya, por tanto, en la variación restante de x_j , es decir, en aquella que no puede explicarse por las demás variables. Siendo R_j^2 el coeficiente de determinación de la regresión de x_j sobre el resto regresores, la varianza del estimador viene dada por

$$\mathbb{V}(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2) \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}, \quad (3.5)$$

Esta expresión muestra que, cuanto mejor pueden las demás variables explicar x_j , mayor es la incertidumbre asociada a $\hat{\beta}_j$. El término $(1 - R_j^2)^{-1}$ recibe el nombre de factor de inflación de la varianza. Cuando R_j^2 se aproxima a uno, significa que x_j es casi completamente reproducible a partir de los demás regresores. En consecuencia, queda muy poca variación propia de x_j con la que identificar su efecto y la varianza de $\hat{\beta}_j$ aumenta considerablemente. Este fenómeno se conoce como colinealidad. No introduce sesgo en los estimadores, pero puede hacer que los coeficientes sean difíciles de estimar con precisión y, por tanto, dificulta interpretar por separado el efecto de variables muy relacionadas entre sí. Esta cuestión reaparecerá al analizar las estimaciones del capítulo 4.

3.1.1.5. Modelos lineales generalizados

Cuando la variable de respuesta no es continua, o cuando su varianza depende de su media, la hipótesis de homocedasticidad de (3.3) deja de ser sostenible. Los modelos lineales generalizados [28, 27] sustituyen (3.2) por tres componentes.

La *componente aleatoria* supone que y_i sigue una distribución de la familia exponencial de dispersión,

$$p(y_i; \theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}, \quad (3.6)$$

donde θ_i es el parámetro canónico, ϕ el parámetro de dispersión y b una función dos veces diferenciable. De (3.6) se deducen

$$\mu_i = \mathbb{E}[y_i] = b'(\theta_i), \quad \mathbb{V}[y_i] = a(\phi) b''(\theta_i) = a(\phi) V(\mu_i), \quad (3.7)$$

siendo V la función de varianza, que caracteriza a la familia. La *componente sistemática* es el predictor lineal $\eta_i = x_i' \beta$. La *función de enlace* g , monótona y diferenciable, relaciona ambas mediante $g(\mu_i) = \eta_i$.

Los dos casos empleados en este trabajo son los siguientes. La familia gaussiana con enlace identidad tiene $V(\mu) = 1$ y reproduce exactamente (3.2). La familia Gamma con enlace logarítmico tiene $V(\mu) = \mu^2$, es decir, desviación típica proporcional a la media y $\mu_i = \exp(x_i' \beta)$, de modo que el efecto de cada regresor es multiplicativo: un incremento unitario de x_j multiplica la media por e^{β_j} . Exige, eso sí, que la respuesta tome valores estrictamente positivos.

La estimación se realiza por máxima verosimilitud. Derivando el logaritmo de (3.6) respecto de β y aplicando la regla de la cadena se obtiene el sistema de ecuaciones de verosimilitud [27,

cap. 2]

$$U(\beta) = \sum_{i=1}^n \frac{y_i - \mu_i}{a(\phi) V(\mu_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} x_i = 0, \quad (3.8)$$

que en general no admite solución cerrada. El método de *scoring* de Fisher lo resuelve iterativamente y equivale a una sucesión de ajustes por mínimos cuadrados ponderados:

$$\beta^{(t+1)} = (X'W^{(t)}X)^{-1}X'W^{(t)}z^{(t)}, \quad (3.9)$$

donde $W^{(t)} = \text{diag}(w_i^{(t)})$, con $w_i = (\partial \mu_i / \partial \eta_i)^2 / (a(\phi)V(\mu_i))$, y $z_i^{(t)} = \eta_i^{(t)} + (y_i - \mu_i^{(t)}) \partial \eta_i / \partial \mu_i$ es la respuesta ajustada. El procedimiento figura en el algoritmo 1.

Algoritmo 1 *Scoring* de Fisher para modelos lineales generalizados

Require: y , X , función de varianza V , enlace g , tolerancia ϵ

```

1:  $\mu^{(0)} \leftarrow y$ ,  $\eta^{(0)} \leftarrow g(\mu^{(0)})$ ,  $t \leftarrow 0$ 
2: repeat
3:    $z_i^{(t)} \leftarrow \eta_i^{(t)} + (y_i - \mu_i^{(t)}) g'(\mu_i^{(t)})$ 
4:    $w_i^{(t)} \leftarrow [g'(\mu_i^{(t)})^2 V(\mu_i^{(t)})]^{-1}$ 
5:    $\beta^{(t+1)} \leftarrow (X'W^{(t)}X)^{-1}X'W^{(t)}z^{(t)}$ 
6:    $\eta^{(t+1)} \leftarrow X\beta^{(t+1)}$ ;  $\mu^{(t+1)} \leftarrow g^{-1}(\eta^{(t+1)})$ 
7:    $t \leftarrow t + 1$ 
8: until  $\|\beta^{(t)} - \beta^{(t-1)}\| < \epsilon$ 
9: return  $\beta^{(t)}$ 
```

Al converger, la matriz $X'\hat{W}X$ es la matriz de información de Fisher del modelo, y su inversa proporciona la matriz de covarianzas asintótica del estimador, $\widehat{V}(\hat{\beta}) = (X'\hat{W}X)^{-1}$, de la que se obtienen los errores estándar de los coeficientes. La bondad del ajuste se mide mediante la *devianza*

$$D = 2[\ell(y; y) - \ell(\hat{\mu}; y)], \quad (3.10)$$

diferencia entre la log-verosimilitud del modelo saturado y la del ajustado, que en el caso gaussiano se reduce a la suma de cuadrados de los residuos.

3.1.1.6. Regresión penalizada

La expresión (3.5) muestra que, en presencia de colinealidad acusada, la varianza del estimador (3.4) puede ser muy elevada, hasta el punto de volverlo inservible. Las penalizaciones atacan ese problema añadiendo al criterio un término que castiga el tamaño de los coeficientes:

$$\hat{\beta}^{\text{ridge}} = \arg \min_b \|y - Xb\|_2^2 + \lambda \|b\|_2^2 = (X'X + \lambda I_{p+1})^{-1}X'y, \quad (3.11)$$

$$\hat{\beta}^{\text{lasso}} = \arg \min_b \|y - Xb\|_2^2 + \lambda \|b\|_1, \quad (3.12)$$

donde $\lambda \geq 0$ regula la intensidad de la penalización. El efecto de (3.11), debida a Hoerl y Kennard [20], se lee con en la descomposición en valores singulares $X = UDV'$, con $D = \text{diag}(d_1, \dots, d_{p+1})$:

$$\hat{y}^{\text{ridge}} = \sum_j u_j \frac{d_j^2}{d_j^2 + \lambda} u_j' y. \quad (3.13)$$

Cada dirección del espacio de regresores queda multiplicada por un factor de contracción $d_j^2/(d_j^2 + \lambda)$, próximo a uno cuando d_j es grande y próximo a cero cuando d_j es pequeño. La contracción es, por tanto, selectiva: afecta sobre todo a las direcciones con poca variación muestral, que son precisamente las responsables de la inestabilidad descrita por (3.5). La penalización ℓ_1 de (3.12) [35] carece de solución cerrada, pero anula exactamente algunos coeficientes, por lo que realiza selección de variables además de contracción.

3.1.1.7. Otras funciones de pérdida

La familia lineal admite además sustituir la pérdida cuadrática por otras. La más relevante aquí es la pérdida absoluta $L(y, f) = |y - f|$, cuyo minimizador poblacional es la mediana condicional y que da lugar a la regresión en desviación absoluta [22]. Se recurrirá a ella en la sección 3.1.2, al comprobar que la elección de la pérdida influye tanto como la del algoritmo cuando la métrica de evaluación es el error absoluto medio.

3.1.2. Modelos de tipo *boosting*

Los modelos de tipo *boosting* empleados en este trabajo se construyen por ensamblaje de árboles de decisión: el predictor final es una suma de muchos árboles pequeños, cada uno ajustado para corregir el error que dejan los anteriores. Conviene por ello describir primero el árbol de decisión y a continuación el mecanismo de ensamblaje.

3.1.2.1. El árbol de regresión

Un árbol de regresión [8] aproxima la función de regresión mediante una función constante a trozos sobre una partición del espacio de regresores en M regiones rectangulares disjuntas $R_1, \dots, R_M$:

$$f(x) = \sum_{m=1}^M c_m \mathbf{1}\{x \in R_m\}, \quad \hat{c}_m = \frac{1}{|R_m|} \sum_{i: x_i \in R_m} y_i. \quad (3.14)$$

Cada región R_m recibe el nombre de *hoja* del árbol. Una hoja es, por tanto, un subconjunto del espacio de regresores delimitado por una conjunción de condiciones del tipo $x_j \leq s$, y $|R_m|$ denota el número de observaciones de la muestra que caen en ella. La predicción del árbol para cualquier punto de una hoja es una constante: la media de las respuestas observadas en ella, que es el valor que minimiza la suma de cuadrados dentro de la región.

La partición no se busca entre todas las posibles, problema de complejidad prohibitiva, sino de forma recursiva y voraz. Partiendo de una única región que contiene toda la muestra, en cada nodo se examina cada regresor j y cada umbral s , y se elige el par que más reduce la suma de cuadrados dentro de los nodos hijos:

$$(j^*, s^*) = \arg \min_{j, s} \left[\min_{c_1} \sum_{x_i \in R_{iz}(j, s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_{de}(j, s)} (y_i - c_2)^2 \right], \quad (3.15)$$

con $R_{iz}(j, s) = \{x : x_j \leq s\}$ y $R_{de}(j, s) = \{x : x_j > s\}$. El procedimiento se repite sobre cada nodo hijo hasta que se cumple alguno de los *criterios de parada*, que son los que impiden que el árbol crezca hasta aislar cada observación en su propia hoja:

- **Profundidad máxima:** número máximo de cortes encadenados desde la raíz hasta una hoja.
- **Número mínimo de observaciones por hoja:** no se abre un corte que deje algún hijo con menos observaciones de las fijadas.
- **Número máximo de hojas:** cota sobre el tamaño total del árbol.
- **Reducción mínima de la pérdida:** no se abre un corte cuya mejora del criterio (3.15) no supere un umbral prefijado.

Dos propiedades del árbol resultan decisivas para lo que sigue. La primera es la varianza elevada del estimador: la estructura de cortes depende de la muestra de forma discontinua, de modo que una perturbación pequeña de los datos puede alterar el primer corte y con él

todo el árbol. La segunda se enuncia como proposición porque justifica una decisión central del capítulo 4.

Proposición 3.3 (Los árboles no extrapolan). *Sea $\hat{f}$ de la forma (3.14), ajustada sobre $\{(x_i, y_i)\}$. Entonces $\hat{f}(x) \in [\min_i y_i, \max_i y_i]$ para todo $x \in \mathbb{R}^p$. La propiedad se conserva bajo combinaciones convexas: si $\hat{F} = \sum_b w_b \hat{f}_b$ con $w_b \geq 0$ y $\sum_b w_b = 1$, entonces $\hat{F}(x)$ pertenece al mismo intervalo.*

Demostración. Cada $\hat{c}_m$ es una media de valores observados de y , luego pertenece al intervalo. Como $\hat{f}(x)$ vale $\hat{c}_m$ para la única hoja que contiene a x , se sigue lo primero. Por otra parte, una combinación convexa de elementos de un intervalo pertenece a ese intervalo, lo que da lo segundo. $\square$

La consecuencia práctica es que un modelo construido con árboles nunca predice valores fuera del recorrido de la respuesta observada en el ajuste, y solo puede emplear un regresor para *partir* el espacio, nunca para *sumar* su valor a la predicción. Un modelo lineal, en cambio, sí puede: si un regresor mantiene con la respuesta una relación aproximadamente lineal, el modelo estima su pendiente y la extiende fuera del rango observado. Esta asimetría es la que determina, en la sección 4.3.2, la definición de la variable a predecir.

3.1.2.2. El compromiso entre sesgo y varianza

Antes de describir el ensamblaje conviene precisar qué se persigue con él. Para un punto x fijo, y suponiendo $y = f(x) + \varepsilon$ con $\mathbb{E}[\varepsilon] = 0$ y $\mathbb{V}[\varepsilon] = \sigma^2$, el error cuadrático esperado del predictor $\hat{f}$ se descompone en

$$\mathbb{E}[(y - \hat{f}(x))^2] = \underbrace{\sigma^2}_{\text{irreducible}} + \underbrace{(\mathbb{E}[\hat{f}(x)] - f(x))^2}_{\text{sesgo}^2} + \underbrace{\mathbb{V}[\hat{f}(x)]}_{\text{varianza}}. \quad (3.16)$$

El primer término no depende del modelo y no puede reducirse. El segundo mide el error sistemático que se comete por imponer una forma funcional demasiado rígida; el tercero, la inestabilidad de las estimaciones ante cambios en la muestra. Interesa que un modelo tenga simultáneamente sesgo bajo y varianza baja, pero ambos objetivos entran habitualmente en conflicto: un modelo lineal tiene sesgo alto y varianza baja, y un árbol crecido sin restricciones, lo contrario.

Las dos grandes familias de ensamblaje atacan cada una un término distinto de (3.16). La agregación *bootstrap* [9] promedia B árboles ajustados sobre remuestras con reemplazamiento de la muestra original, y reduce la *varianza*: si las estimaciones realizadas por cada uno de los B árboles tienen varianza σ_a^2 y la correlación entre dos cualesquiera de ellas es ρ , la varianza del promedio vale

$$\mathbb{V}\left[\frac{1}{B} \sum_{b=1}^B \hat{f}_b(x)\right] = \frac{1}{B^2} [B\sigma_a^2 + B(B-1)\rho\sigma_a^2] = \rho\sigma_a^2 + \frac{1-\rho}{B}\sigma_a^2, \quad (3.17)$$

expresión cuyo segundo sumando se anula al crecer B mientras que el primero permanece. El *boosting*, en cambio, reduce el *sesgo*, ajustando los árboles en cadena de modo que cada uno corrija lo que los anteriores no explicaron. Es esta segunda vía la que se emplea aquí; la conveniencia de reducir además el término $\rho\sigma_a^2$ de (3.17) es la que justifica el submuestreo aleatorio que se describe más abajo.

3.1.2.3. *Boosting* por gradiente

El *boosting* por gradiente [15] construye el predictor de forma aditiva realizando un descenso de gradiente en el espacio de funciones. Partiendo de una predicción constante, en la iteración m se calculan los *seudoresiduos*, que son las derivadas de la función de pérdida respecto de la predicción, cambiadas de signo y evaluadas en el modelo vigente:

$$r_{im} = - \left. \frac{\partial L(y_i, F)}{\partial F} \right|_{F=F_{m-1}(x_i)}. \quad (3.18)$$

Cada r_{im} indica en qué dirección y con qué intensidad convendría mover la predicción de la observación i para reducir la pérdida. Se ajusta a continuación un árbol que aproxime ese conjunto de direcciones y se actualiza el modelo avanzando un paso pequeño en ella:

$$F_m(x) = F_{m-1}(x) + \eta \sum_{j=1}^J \gamma_{jm} \mathbf{1}\{x \in R_{jm}\}, \quad \gamma_{jm} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma), \quad (3.19)$$

donde $\eta \in (0, 1]$ es la tasa de aprendizaje y $R_{1m}, \dots, R_{Jm}$ son las hojas del árbol ajustado en esa iteración.

Con pérdida cuadrática, $L(y, F) = \frac{1}{2}(y - F)^2$, se tiene $r_{im} = y_i - F_{m-1}(x_i)$, es decir, el residuo ordinario. Con pérdida absoluta, $L(y, F) = |y - F|$, se tiene $r_{im} = \text{sign}(y_i - F_{m-1}(x_i))$: el árbol aprende únicamente el signo del error y no su magnitud, lo que confiere al procedimiento robustez frente a observaciones atípicas. El procedimiento completo figura en el algoritmo 2.

Algoritmo 2 *Boosting* por gradiente con árboles de regresión

Require: muestra $\{(x_i, y_i)\}_{i=1}^n$; pérdida L ; número de iteraciones M ; tasa de aprendizaje η ; número de hojas J

```

1:  $F_0(x) \leftarrow \arg \min_c \sum_{i=1}^n L(y_i, c)$ 
2: for  $m = 1, \dots, M$  do
3:   for  $i = 1, \dots, n$  do
4:      $r_{im} \leftarrow -[\partial L(y_i, F)/\partial F]_{F=F_{m-1}(x_i)}$ 
5:   end for
6:   ajustar un árbol de  $J$  hojas  $R_{1m}, \dots, R_{Jm}$  a los pares  $\{(x_i, r_{im})\}$ 
7:   for  $j = 1, \dots, J$  do
8:      $\gamma_{jm} \leftarrow \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)$ 
9:   end for
10:   $F_m(x) \leftarrow F_{m-1}(x) + \eta \sum_{j=1}^J \gamma_{jm} \mathbf{1}\{x \in R_{jm}\}$ 
11: end for
12: return  $\hat{f} = F_M$ 

```

3.1.2.4. LightGBM

La implementación empleada en este trabajo es LightGBM [21], que introduce sobre el esquema del algoritmo 2 tres modificaciones.

La primera es el uso de una aproximación de segundo orden de la pérdida, introducida por XGBoost [11]. Escribiendo $g_i = \partial_F L(y_i, F_{m-1}(x_i))$ y $h_i = \partial_F^2 L(y_i, F_{m-1}(x_i))$, y añadiendo una penalización sobre los valores de las hojas, el objetivo de cada iteración queda

$$\tilde{\mathcal{L}}^{(m)} = \sum_{i=1}^n \left[g_i f_m(x_i) + \frac{1}{2} h_i f_m(x_i)^2 \right] + \frac{\lambda}{2} \sum_{j=1}^J w_j^2, \quad (3.20)$$

que es una función cuadrática de los valores de hoja w_j . Minimizándola en cada hoja I_j se obtiene, en forma cerrada, el valor óptimo y la ganancia asociada a un corte que divide un nodo en dos hijos:

$$w_j^* = -\frac{G_j}{H_j + \lambda}, \quad \text{ganancia} = \frac{1}{2} \left[\frac{G_{iz}^2}{H_{iz} + \lambda} + \frac{G_{de}^2}{H_{de} + \lambda} - \frac{(G_{iz} + G_{de})^2}{H_{iz} + H_{de} + \lambda} \right], \quad (3.21)$$

con $G_j = \sum_{i \in I_j} g_i$ y $H_j = \sum_{i \in I_j} h_i$. La segunda expresión de (3.21) sustituye a (3.15) como criterio de corte, de modo que la penalización λ interviene en la propia construcción de la estructura del árbol y no solo en el valor de sus hojas. Nótese además que w_j^* es una media de gradientes contraída hacia cero, con λ desempeñando el papel de constante de contracción.

La segunda modificación es la discretización previa de cada regresor en un histograma de K intervalos, lo que reduce el coste de buscar el mejor corte de $\mathcal{O}(n)$ a $\mathcal{O}(K)$ por variable, con K típicamente igual a 255.

La tercera es el crecimiento *por hojas*: en lugar de expandir todos los nodos de un nivel antes de pasar al siguiente, en cada paso se expande únicamente la hoja que promete mayor ganancia según (3.21). Esto produce árboles más profundos y expresivos con el mismo número de hojas y, por ello, con mayor propensión al sobreajuste, lo que obliga a controlar simultáneamente el número de hojas y el tamaño mínimo de hoja. El algoritmo 3 recoge el procedimiento de construcción de un árbol.

Algoritmo 3 Construcción de un árbol en LightGBM

Require: gradientes $\{g_i\}$ y hessianos $\{h_i\}$; número de hojas J ; tamaño mínimo de hoja $n_{\min}$; penalización λ ; histogramas de K intervalos

- 1: iniciar el árbol con una única hoja que contiene todas las observaciones
 - 2: **while** el número de hojas sea menor que J **do**
 - 3: **for** cada hoja I del árbol aún no evaluada **do**
 - 4: acumular (g_i, h_i) en los K intervalos del histograma de cada regresor
 - 5: evaluar la ganancia (3.21) de cada corte candidato y retener el mejor de entre los que dejan al menos $n_{\min}$ observaciones en cada hijo
 - 6: **end for**
 - 7: **if** ninguna ganancia es positiva, **break**
 - 8: expandir la hoja cuyo mejor corte tenga la ganancia máxima
 - 9: **end while**
 - 10: asignar a cada hoja I_j el valor w_j^* de (3.21)
 - 11: **return** el árbol
-

3.1.2.5. Hiperparámetros

Se denomina *hiperparámetro* a toda cantidad que gobierna el procedimiento de ajuste y que, a diferencia de los parámetros β de un modelo lineal, no se obtiene resolviendo (3.1) sobre la muestra de ajuste. La razón es que el criterio (3.1) es monótono en la flexibilidad del modelo: entendiendo por modelo *flexible* aquel capaz de representar un conjunto más amplio de funciones —un árbol con más hojas, un ensamblaje con más iteraciones—, un modelo más flexible alcanza siempre un error de ajuste menor o igual. Buscar los hiperparámetros que minimizan (3.1) conduciría por tanto al modelo más flexible posible y, con él, al sobreajuste. Deben elegirse evaluando el modelo sobre datos que no hayan intervenido en su ajuste, según se detalla en la sección 3.2.8.

Los del modelo empleado son ocho, y conviene enunciarlos junto con el mecanismo por el que cada uno controla la complejidad. El **número de árboles** M es el primero: cada árbol reduce sesgo, de modo que existe un valor óptimo finito, ya que a partir de él las

iteraciones adicionales ajustan ruido. Va acompañado de la **tasa de aprendizaje** η , que escala la actualización (3.19); valores pequeños exigen más iteraciones pero generalizan mejor, y como el producto ηM gobierna la capacidad total, ambos deben ajustarse conjuntamente. El **número de hojas** J limita el orden de las interacciones representables, puesto que un árbol equilibrado con J hojas no puede expresar interacciones de más de $\log_2 J$ variables. El **tamaño mínimo de hoja** actúa como cota inferior a $|I_j|$ e impide hojas sostenidas por muy pocas observaciones, cuya media es esencialmente ruido. Dos hiperparámetros introducen aleatoriedad: el **submuestreo de filas** ajusta cada árbol sobre una fracción aleatoria de la muestra, lo que reduce ρ en (3.17), y el **submuestreo de columnas** restringe cada corte a un subconjunto aleatorio de regresores, con el mismo efecto e impidiendo además que un regresor dominante monopolice la estructura. La **penalización** λ aparece en (3.20) y (3.21): contrae los valores de hoja y penaliza los cortes sostenidos por poca masa. Y la **función de pérdida** L determina qué característica de la distribución condicional se estima, la media si es cuadrática y la mediana si es absoluta.

3.1.2.6. Interpretación: valores de Shapley

Los modelos construidos por ensamblaje de árboles de decisión no permiten una interpretación directa de sus predicciones, como sí ocurre en los modelos lineales a través de sus parámetros, o en un árbol individual a través del camino de cortes que conduce a cada hoja. Los valores de Shapley [34, 25] proporcionan una descomposición aditiva de cada predicción que suple esa carencia.

Dado un conjunto N de p jugadores y una función característica $v : 2^N \rightarrow \mathbb{R}$ que asigna a cada coalición la ganancia que obtiene, el valor de Shapley del jugador j es

$$\varphi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(p - |S| - 1)!}{p!} [v(S \cup \{j\}) - v(S)], \quad (3.22)$$

es decir, el promedio de su contribución marginal sobre todos los órdenes posibles de formación de la coalición. Se demuestra que (3.22) es el único reparto que satisface simultáneamente eficiencia ($\sum_j \varphi_j = v(N) - v(\emptyset)$), simetría, nulidad y aditividad.

Tomando como jugadores las variables explicativas y $v(S) = \mathbb{E}[\hat{f}(x) \mid x_S]$ se obtiene

$$\hat{f}(x) = \varphi_0 + \sum_{j=1}^p \varphi_j(x), \quad \varphi_0 = \mathbb{E}[\hat{f}(x)], \quad (3.23)$$

donde $\varphi_j(x)$ es la contribución de la variable j a esa predicción concreta, expresada en las unidades de la respuesta. La importancia global de la variable j se define como el promedio $\frac{1}{n} \sum_i |\varphi_j(x_i)|$. El cálculo exacto de (3.22) requiere recorrer 2^{p-1} subconjuntos; para conjuntos de árboles existe, sin embargo, un algoritmo exacto de coste polinómico [26].

Dos limitaciones deben declararse. Las contribuciones explican el modelo y no el fenómeno: bajo colinealidad, el reparto entre dos variables depende de cuál haya utilizado el algoritmo y no de cuál sea la causa. Y la aditividad de (3.23) no implica que el efecto de las variables sea aditivo, sino únicamente que lo es la atribución.

3.1.3. Predictores de referencia

Para juzgar si un modelo aporta algo es imprescindible compararlo con predictores deliberadamente sencillos, que acotan por abajo el desempeño alcanzable. Se emplean tres. El *predictor constante* $\hat{f}(x) \equiv c$ ignora las variables explicativas; la solución de (3.1) es la media muestral si L es cuadrática y la mediana muestral si es absoluta, y es además la referencia

implícita del coeficiente de determinación. El *predictor constante por grupo* estima una constante distinta dentro de cada grupo definido por las variables categóricas más relevantes; es una referencia más exigente, pero su varianza crece al reducirse el número de observaciones por grupo y solo puede aplicarse a grupos observados durante el ajuste. El *predictor de persistencia* supone que la magnitud de interés no cambiará respecto de su último valor observado; es la referencia natural cuando la respuesta es un incremento, pues equivale a predecir un incremento nulo.

3.2. Metodología

3.2.1. Planteamiento

Esta sección expone la metodología seguida sin referencia al problema de aplicación, de modo que pueda juzgarse por sí misma; el capítulo 4 la recorre después sobre los datos concretos.

El objeto es construir, a partir de datos ya observados, una función que en un instante determinado prediga una magnitud que solo se conocerá más tarde, empleando exclusivamente la información disponible en ese instante. Enunciado así, el problema encierra dos decisiones que gobiernan todas las demás.

La primera es *qué* se predice exactamente. Una misma magnitud admite definiciones alternativas de la variable de respuesta, y la elección no es accesoria: predecir el nivel de una serie o predecir su incremento entre dos instantes son problemas distintos, con recorridos distintos y dificultad distinta. Cuando el nivel varía mucho entre unidades y poco dentro de cada una, predecirlo directamente obliga al modelo a reconstruir un nivel que las variables explicativas apenas identifican, mientras que predecir el incremento lo elimina por diferencia y concentra el esfuerzo en la parte del fenómeno que verdaderamente interesa. Además, según se ha visto en la sección 3.1, esa elección afecta de manera desigual a cada familia de modelos: es indiferente para un modelo lineal que incluya el nivel de partida entre sus regresores (proposición 3.1) y decisiva para uno construido con árboles (proposición 3.3). La metodología contempla por ello comparar varias definiciones del objetivo en lugar de fijar una de antemano.

La segunda es *cómo* se comprueba que lo aprendido sirve para datos nuevos. Un error medido sobre los mismos datos con los que se ha ajustado el modelo no informa sobre su comportamiento futuro, y la partición de la muestra que corrige ese sesgo debe quedar fijada antes de ajustar nada, no elegirse a la vista de los resultados.

3.2.2. Metodologías de modelización predictiva

La construcción de un modelo predictivo a partir de datos observados no es una sucesión de decisiones improvisadas, sino un proceso estructurado en fases para el que existen marcos metodológicos estándar. Dos de ellos se han consolidado en la práctica profesional.

El primero es **CRISP-DM** (*Cross-Industry Standard Process for Data Mining*), publicado por un consorcio industrial en el que participó SPSS —posteriormente integrada en IBM— y documentado en una guía de referencia [10]. Se articula en seis fases: comprensión del problema, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Su rasgo distintivo es que no es lineal: la guía prevé de forma explícita el retorno a fases anteriores, porque lo que se descubre al explorar los datos obliga con frecuencia a reformular el problema, y lo que se observa al evaluar un modelo obliga a rehacer la preparación.

El segundo es **SEMMA** (*Sample, Explore, Modify, Model, Assess*), propuesto por SAS Institute como estructura de trabajo de su herramienta Enterprise Miner [32]. Consta de cinco etapas: muestreo, exploración, modificación, modelado y evaluación. A diferencia de CRISP-DM, deja fuera tanto la fase inicial de comprensión del problema como la final de despliegue,

y se ciñe al ciclo estrictamente analítico.

La comparación entre ambos [3] muestra que las cinco etapas de SEMMA se corresponden con las fases centrales de CRISP-DM, de modo que la diferencia real no está en el trabajo estadístico que prescriben —que es el mismo— sino en su alcance: CRISP-DM abarca el ciclo completo de un proyecto y SEMMA únicamente su núcleo analítico.

La metodología seguida en este trabajo se ajusta a ese esquema general. Al tratarse de un trabajo de investigación y no de un desarrollo destinado a producción, no contempla fase de despliegue, y la comprensión del problema se ha tratado por separado en el capítulo 2. Las fases restantes se organizan como sigue:

- **Construcción de la tabla de entrada e imposición de filtros**, que define la población de análisis.
- **Definición de la variable de respuesta y generación de variables explicativas**.
- **Análisis exploratorio** de las variables construidas.
- **Tratamiento de variables**: valores ausentes, valores atípicos y transformaciones.
- **Ajuste de los modelos** y selección de sus hiperparámetros.
- **Validación e interpretación** de los resultados.

Las dos primeras fases fijan qué se predice; la tercera y la cuarta, con qué; la quinta y la sexta, cómo se ajusta y cómo se comprueba que lo aprendido sirve para datos nuevos. La figura 3.1 las representa en el orden en que se recorren, junto con el retorno a la primera fase que ambas metodologías prescriben y que este trabajo ha necesitado en varias ocasiones.

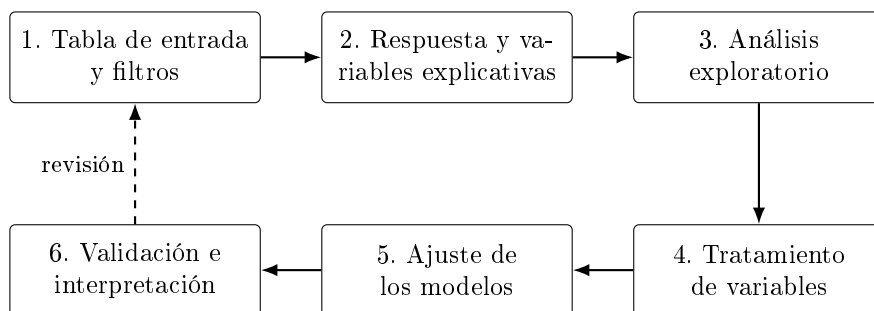


FIGURA 3.1: Fases de la metodología seguida y orden en que se recorren. La flecha discontinua representa el retorno a fases anteriores que CRISP-DM [10] prescribe de forma explícita.

3.2.3. Generalización y partición de la muestra

Antes de describir las fases conviene fijar el criterio con el que se juzgará todo lo que en ellas se decida, porque condiciona la construcción de la tabla desde el primer momento.

El objetivo de un modelo predictivo no es reproducir los datos con los que se ha ajustado, sino generalizar a datos nuevos. Como el criterio (3.1) es monótono en la flexibilidad del modelo, el error medido sobre la propia muestra de ajuste es un estimador optimista del error de generalización, y tanto más cuanto más flexible sea el modelo. De ahí la necesidad de dividir la tabla de entrada en al menos dos muestras disjuntas:

- La muestra de **entrenamiento**, con la que se estiman los parámetros del modelo.
- La muestra de **prueba** o *test*, que no interviene en ninguna decisión y sobre la que se mide el error que se reporta como estimación del error de generalización.

Cuando además hay que elegir hiperparámetros, esa división resulta insuficiente. Elegirlos según el error sobre la muestra de prueba la convertiría en parte del procedimiento de ajuste, y su error dejaría de ser una estimación válida: se estaría empleando dos veces, una para decidir y otra para medir. Conviene por ello contemplar una tercera muestra:

- La muestra de **validación**, empleada exclusivamente para comparar configuraciones de hiperparámetros y quedarse con la mejor.

El esquema completo consiste, por tanto, en estimar los parámetros sobre entrenamiento, elegir los hiperparámetros según el error sobre validación y reportar, como medida final, el error sobre prueba.

Cuando las observaciones están ordenadas en el tiempo, las tres muestras deben respetar además ese orden: entrenamiento primero, validación después y prueba al final. El motivo es que una partición aleatoria asignaría al entrenamiento observaciones posteriores a las de prueba, de modo que el modelo dispondría de información que en uso real no existe todavía; por ejemplo, podría aprender de una carrera de octubre para predecir una de julio. El error estimado sería entonces optimista, y lo sería sin que nada en el experimento lo delatara: el ajuste converge igual, las métricas salen mejores y no se produce ningún aviso [5].

Esta exigencia no es una cautela adicional que se aplique al final, sino una restricción que atraviesa todas las fases: cualquier cantidad estimada a partir de los datos —una mediana para imputar, un umbral de recorte, la codificación de una variable categórica— debe calcularse únicamente sobre el tramo de ajuste, o la muestra de prueba dejará de ser independiente del procedimiento que se pretende evaluar.

3.2.4. Construcción de la tabla de entrada e imposición de filtros

La primera decisión es la *unidad de observación*: qué representa una fila de la tabla. Cuando los datos tienen estructura anidada —observaciones agrupadas en unidades— conviene distinguir dos niveles: el de la unidad, que define la población de estudio, y el de la observación, que define la muestra de ajuste. Ambos deben declararse, porque el segundo determina el tamaño muestral aparente y el primero el efectivo.

Los filtros restringen la muestra a las observaciones sobre las que el modelo va a operar. Dos exigencias los gobiernan. Primera, todo filtro debe poder aplicarse en el momento de predecir; uno que dependa de información posterior invalidaría la evaluación. Segunda, si un filtro condiciona sobre la propia variable de respuesta —descartar observaciones por su valor— las conclusiones se refieren a la subpoblación filtrada y no a la original, lo que debe declararse antes de presentar resultado alguno.

3.2.5. Definición de la variable de respuesta y generación de variables explicativas

Sea $\{y_{it}\}$ una colección de observaciones agrupadas en unidades i e indexadas por t dentro de cada unidad. Se consideran dos formas de definir la variable a predecir.

La primera es predecir directamente el *nivel* y_{it} . La segunda es predecir su *incremento* a un horizonte h fijo,

$$\Delta_{it}^{(h)} = y_{i,t+h} - y_{i,t}, \quad (3.24)$$

empleando exclusivamente información disponible hasta el instante t .

La elección entre ambas se relaciona directamente con la proposición 3.3. Cuando el nivel de la respuesta varía mucho entre unidades, un modelo construido con árboles debe reconstruir ese nivel particionando el espacio, y no puede predecir niveles que no haya observado. Si en cambio se predice un incremento, el nivel propio de cada unidad desaparece por diferencia y el

objetivo pasa a moverse en un recorrido estrecho y común a todas las unidades. Una variante intermedia consiste en restar al nivel un estimador $\hat{a}_i$ construido con información previa de la propia unidad y disponible en el momento de predecir; esta transformación se denominará *anclaje* y $\hat{a}_i$ recibirá el nombre de *ancla*.

Las dos proposiciones enunciadas en la sección 3.1 delimitan qué cabe esperar de cada familia de modelos ante esta elección. Por la proposición 3.1, un modelo lineal que incluya el ancla entre sus regresores predice exactamente lo mismo con el nivel que con el objetivo anclado: el anclaje es para él una operación inocua. Por la proposición 3.3, un modelo de árboles solo puede emplear el ancla para partir el espacio, nunca para sumarla a la predicción; anclar el objetivo traslada esa suma a la definición de la variable a predecir y libera al árbol de tener que representarla. Cabe esperar, por tanto, que el anclaje deje indiferente al modelo lineal y mejore apreciablemente al de árboles, y que la comparación entre ambas definiciones sea informativa únicamente para el segundo.

La *generación de variables explicativas* consiste en construir, a partir de los datos brutos, regresores con contenido informativo sobre la respuesta. Se distinguen tres clases: variables observadas directamente; variables derivadas, obtenidas por agregación o transformación dentro de la unidad; y codificaciones numéricas de variables categóricas de muchos niveles. Estas últimas, cuando se construyen con el valor medio de la respuesta en cada nivel, deben contraerse hacia la media global mediante

$$\hat{\theta}_g^{\text{enc}} = w_g \hat{\theta}_g + (1 - w_g) \bar{y}, \quad w_g = \frac{n_g}{n_g + K}, \quad (3.25)$$

donde $\hat{\theta}_g$ es la media de la respuesta en el nivel g , n_g su número de observaciones y K una constante de contracción [13, 17]. La expresión (3.25) conserva la estimación propia de los niveles bien representados y aproxima a la media global la de aquellos con pocas observaciones, cuya estimación sería inestable. Debe calcularse exclusivamente sobre el tramo de ajuste.

La restricción que gobierna toda esta fase se enuncia en una línea: toda variable explicativa debe ser calculable con la información disponible en el instante en que se emite la predicción. Su verificación es una cuestión de contabilidad temporal, no estadística, y su incumplimiento no produce ningún síntoma detectable en las métricas.

3.2.6. Análisis exploratorio

El análisis exploratorio persigue dos fines: caracterizar la distribución de cada variable y contrastar, antes de ajustar ningún modelo, si las relaciones esperadas se observan en los datos.

El descriptivo univariante reporta, para cada variable, media, desviación típica, mínimo, cuartiles, máximo y los coeficientes de asimetría y curtosis,

$$\gamma_1 = \frac{\mathbb{E}[(x - \mu)^3]}{\sigma^3}, \quad \gamma_2 = \frac{\mathbb{E}[(x - \mu)^4]}{\sigma^4} - 3, \quad (3.26)$$

que valen ambos cero bajo normalidad y cuya magnitud anticipa la presencia de observaciones extremas.

El descriptivo bivalente se realiza mediante *medias por tramo*. Dada una variable x , se definen B tramos a partir de sus deciles muestrales y se calcula, en cada tramo b ,

$$\bar{y}_b = \frac{1}{n_b} \sum_{i \in b} y_i, \quad \widehat{\text{e}}(\bar{y}_b) = \frac{s_b}{\sqrt{n_b}}, \quad (3.27)$$

representando $\bar{y}_b \pm 1,96 \widehat{\text{e}}(\bar{y}_b)$ frente al centro del tramo. Se prefiere a la correlación de Pearson por dos motivos: no presupone linealidad, de modo que revela relaciones no monótonas que un coeficiente promediaría a cero; y cuantifica la incertidumbre de cada punto, evitando leer como tendencia lo que es ruido de un tramo poco poblado.

3.2.7. Tratamiento de variables

Valores ausentes. Su tratamiento depende del mecanismo que los genera. Si la ausencia es independiente del valor no observado y de la respuesta, cualquier imputación razonable es admisible. Si la ausencia es informativa, imputar destruye información, y es preferible codificarla explícitamente mediante una variable indicadora que acompañe a la imputación. Los modelos de la forma (3.14) admiten un tratamiento nativo: la observación ausente se envía al lado del corte que minimiza la pérdida.

Valores atípicos. Se emplean dos reglas de detección. La regla del rango intercuartílico marca como atípica toda observación fuera del intervalo

$$[Q_1 - 1,5(Q_3 - Q_1), Q_3 + 1,5(Q_3 - Q_1)], \quad (3.28)$$

y la regla de las tres desviaciones típicas, toda observación con $|x - \bar{x}| > 3s$. Ambas presuponen simetría aproximada, de modo que su discrepancia sobre una misma variable es indicio de asimetría y no de error de medida. Nótese que la primera exige poder ordenar los valores de la variable para calcular sus cuartiles, por lo que solo tiene sentido sobre variables cuantitativas u ordinales; y que ninguna de las dos es aplicable a variables discretas de pocos niveles, pues si $Q_1 = Q_3$ el rango intercuartílico es nulo y (3.28) marca como atípico todo valor distinto de la mediana.

La decisión sobre qué hacer con ellos debe justificarse. Se consideran tres tratamientos, además de no aplicar ninguno: el recorte o *winsorización* a los percentiles 1 y 99, que conserva la observación pero acota su influencia; la eliminación de las observaciones que (3.28) marca sobre la respuesta; y la transformación logarítmica de las variables de asimetría acusada, que comprime la cola derecha sin descartar información. Cualquiera de ellos se aplica únicamente sobre la muestra de ajuste: la de prueba debe permanecer intacta, pues representa los datos tal como llegarían en uso real.

El resultado que ordena la elección es el siguiente. La influencia de una observación sobre el ajuste depende de la pérdida: bajo pérdida cuadrática, la derivada $\partial L / \partial f = -(y - f)$ crece linealmente con el error, de modo que una observación extrema ejerce una influencia proporcional a su desviación; bajo pérdida absoluta, $\partial L / \partial f = -\text{sign}(y - f)$ está acotada y toda observación ejerce la misma influencia con independencia de su magnitud. En consecuencia, el tratamiento de atípicos y la elección de una pérdida robusta son dos soluciones al mismo problema, y cabe esperar que la segunda haga innecesaria a la primera.

Transformaciones. Las transformaciones monótonas de los regresores son irrelevantes para los modelos de la forma (3.14), puesto que (3.15) depende únicamente del orden de los valores y no de su magnitud. Para los modelos lineales sí importan: la penalización (3.11) no es invariante frente a cambios de escala, lo que obliga a tipificar los regresores para que la contracción afecte por igual a todos los coeficientes.

3.2.8. Ajuste de los modelos

Fijado el esquema de tres muestras de la sección 3.2.3, resta describir cómo se recorre el espacio de hiperparámetros que la muestra de validación permite comparar.

3.2.8.1. Búsqueda de hiperparámetros

La búsqueda exhaustiva sobre una rejilla es inviable cuando hay varios hiperparámetros continuos, y la búsqueda aleatoria [7], aunque eficaz, no aprovecha las evaluaciones previas. La optimización bayesiana mediante el estimador de Parzen [6] modela la densidad de las

configuraciones ya probadas separándolas en dos grupos según su error: $\ell(\theta)$ para aquellas cuyo error queda por debajo de un cuantil dado y $g(\theta)$ para el resto. En cada paso propone la configuración

$$\theta^{(t+1)} = \arg \max_{\theta} \frac{\ell(\theta)}{g(\theta)}, \quad (3.29)$$

esto es, la más verosímil entre las buenas y menos verosímil entre las malas, lo que equivale a maximizar la mejora esperada. Su implementación [1] admite espacios de búsqueda mixtos y escalas logarítmicas.

La partición sobre la que se comparan las configuraciones debe respetar el orden temporal por el motivo expuesto en la sección 3.2.3: los parámetros se estiman sobre el tramo más antiguo y las configuraciones se comparan sobre el siguiente, de modo que la muestra de prueba no interviene en ningún momento de la elección.

3.2.9. Validación e interpretación de resultados

Capacidad predictiva. Se reportan

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad \text{RMSE} = \left(\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right)^{1/2}, \quad R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}. \quad (3.30)$$

Las dos primeras se expresan en las unidades de la respuesta y las minimizan predictores distintos: la mediana condicional el MAE, la media condicional el RMSE. Reportar ambas es informativo precisamente cuando discrepan, pues un modelo con buen MAE y mal RMSE acierta casi siempre y falla mucho ocasionalmente.

Sobre el coeficiente de determinación conviene hacer dos precisiones. La primera es que R^2 compara el modelo con el predictor constante $\bar{y}$, referencia que resulta engañosa cuando la respuesta presenta mucha variación entre unidades y poca dentro de ellas: en ese caso R^2 se aproxima a uno sin que el modelo haya explicado nada de la variación intra-unidad. La segunda es que, calculado sobre una muestra de prueba, R^2 puede tomar valores negativos. Ello ocurre cuando la suma de cuadrados de los residuos del modelo supera a la de las desviaciones respecto de $\bar{y}$, siendo aquí $\bar{y}$ la media de la muestra de prueba, que el modelo no conoce. Un valor negativo indica sencillamente que el modelo predice peor que esa media; no es un error de cálculo, y no lo corrige el coeficiente ajustado, que penaliza por el número de regresores y está pensado para comparar modelos sobre la muestra de ajuste, no para medir generalización.

Todas las métricas se calculan sobre el conjunto de ajuste y sobre el de prueba. La diferencia entre ambos cuantifica el sobreajuste; se resume mediante el cociente $\text{MAE}_{\text{prueba}}/\text{MAE}_{\text{ajuste}}$, que vale uno para un modelo sin capacidad de memorización y crece con ella. Un cociente elevado no descalifica por sí solo a un modelo —la decisión se toma únicamente con el error de prueba— pero indica que el margen de mejora está en regularizar y no en aumentar la flexibilidad.

Capacidad explicativa. Para los modelos lineales generalizados se reporta la tabla de estimaciones $\hat{\beta}_j$ con su error estándar $\widehat{\text{ee}}(\hat{\beta}_j)$ y el estadístico $z_j = \hat{\beta}_j/\widehat{\text{ee}}(\hat{\beta}_j)$, tipificando previamente los regresores para que las magnitudes sean comparables entre sí. Para los modelos ensamblados se reportan los valores de Shapley (3.23), tanto en importancia global como en gráficos de dependencia $x_{ij} \mapsto \varphi_j(x_i)$, que revelan la forma funcional aprendida sin imponerla.

La validación de la capacidad explicativa consiste en contrastar ambas fuentes con el análisis bivalente de la sección 3.2.6: un modelo cuyas atribuciones reproducen relaciones ya visibles en las medias por tramo merece crédito; uno cuyas atribuciones no se corresponden con ninguna relación observable exige explicación.

Capítulo 4

Aplicación de la metodología

Se aplica en este capítulo la metodología de la sección 3.2 al problema descrito en el capítulo 2, cuyo vocabulario técnico —*stint*, compuesto, degradación, aire sucio, *undercut*, pérdida de boxes— se emplea aquí sin volver a definirlo; la tabla 2.1 lo recoge.

4.1. Fuente de datos

Los datos proceden de OpenF1 [29], servicio de acceso libre que expone mediante una interfaz de programación el cronometraje oficial que la Federación Internacional del Automóvil publica en cada sesión. Se ha preferido a las alternativas por cubrir las temporadas completas desde 2023, no exigir credenciales y publicar el detalle vuelta a vuelta. La biblioteca FastF1 [14] se emplea únicamente como fuente de contraste para verificar la coherencia de la información de compuestos.

Se descargan seis archivos. El de reuniones identifica cada Gran Premio; el de sesiones distingue entrenamientos libres, clasificación y carrera; el de vueltas contiene, para cada piloto y vuelta, la duración total y la de cada sector y las indicadoras de entrada y salida de boxes; el de *stints* registra el compuesto montado, la vuelta de inicio, la de fin y la edad del juego; el de intervalos ofrece la diferencia de cada coche con el líder y con el precedente; y el de meteorología registra temperatura del aire y del asfalto, humedad, presión y viento. Se añaden dos tablas construidas por el autor: el calendario real de las tres temporadas y una tabla de características físicas de cada circuito.

El volumen bruto asciende a algo más de dos millones de registros de vuelta y de intervalo correspondientes a las temporadas 2023, 2024 y 2025. Una vez descargados y emparejados los seis recursos, la información se consolida en un único fichero comprimido, que junto con las tablas auxiliares del calendario y de los circuitos constituye la entrada de todo el procesamiento posterior.

Una advertencia sobre la fuente debe hacerse antes de cualquier otra cosa, porque condiciona la partición temporal. El identificador de reunión no respeta el orden del calendario: la reunión de Silverstone de 2025 lleva una clave numéricamente posterior a la de carreras celebradas en otoño. Ordenar por esa clave habría asignado al ajuste observaciones posteriores a las de prueba. Todas las particiones se realizan, por ello, sobre la fecha real de cada Gran Premio.

4.2. Software utilizado

El desarrollo se ha realizado íntegramente en Python 3.11.

Se han empleado NumPy 1.26 para el álgebra matricial y la factorización QR con pivoteo del filtro de rango; pandas 2.2 para el emparejamiento de los seis recursos y la construcción de

la tabla; statsmodels 0.14 para los modelos lineales generalizados por *scoring* de Fisher (3.9), con las familias gaussiana y Gamma; LightGBM 4.7 para el *boosting* por gradiente (3.19), con pérdidas cuadrática y absoluta; Optuna 4.6 para la optimización bayesiana con el estimador de Parzen (3.29); SHAP 0.51 para el cálculo exacto de los valores de Shapley sobre conjuntos de árboles (3.22); SciPy 1.16 para las utilidades numéricas; y Matplotlib 3.9 para las figuras.

El código se organiza en cinco módulos con diferentes finalidades: construcción de la tabla, catálogo de modelos y tratamiento de atípicos, experimento principal, análisis exploratorio e interpretación. Todas las cifras y tablas de este capítulo se generan automáticamente a partir de los ficheros de resultados, sin transcripción manual.

4.3. Aplicación de la metodología

Los apartados que siguen recorren, en el mismo orden, las seis fases de la metodología descrita en la sección 3.2: construcción de la tabla de entrada e imposición de filtros, definición de la variable de respuesta y generación de variables explicativas, análisis exploratorio, tratamiento de variables, ajuste de los modelos y validación e interpretación de los resultados.

4.3.1. Construcción de la tabla de entrada e imposición de filtros

4.3.1.1. Estructura de la tabla

La tabla se organiza en dos niveles, conforme a la sección 3.2.4. El nivel superior es la *unidad*, definida como el par piloto–carrera: un piloto concreto en un Gran Premio concreto, lo que equivale a la terna año–circuito–piloto. El nivel inferior es la *observación*, definida como una vuelta del segundo *stint* de esa unidad. Cada unidad aporta, por tanto, tantas filas como vueltas tenga su segundo *stint*.

Esta elección de unidad no es arbitraria. Dentro de un par piloto–carrera permanecen constantes el coche, el piloto, el trazado, la fecha, la puesta a punto y las condiciones ambientales; lo único que varía entre observaciones es la vuelta que avanza, el neumático que envejece, el combustible que se consume y el tráfico que aparece o desaparece. Es, por tanto, el bloque más fino que conserva varias observaciones en su interior, que es exactamente la propiedad que exige la definición del objetivo de la sección 3.2.5.

4.3.1.2. Filtros

Se aplican cuatro filtros sucesivos, cada uno con su justificación.

El primero restringe a la sesión de carrera y a las vueltas en bandera verde, descartando además las completadas tras un coche de seguridad, las vueltas de entrada y salida de boxes y las tres primeras de cada carrera. El motivo es que esas vueltas contienen variación de tiempo que no procede del coche sino de la neutralización, y que además está correlacionada con la posición. Se descartan también las vueltas que superan el ciento diez por ciento del mejor ritmo de esa carrera. El umbral se justifica porque una vuelta un diez por ciento más lenta que la mejor de la carrera —entre siete y once segundos, según el circuito— no puede explicarse por degradación ni por combustible, magnitudes cuyo efecto conjunto es del orden de un segundo: responde necesariamente a una incidencia, un despiste o una neutralización que el resto de filtros no ha capturado. Es el criterio habitual en la literatura de análisis de carrera [19] y coincide además con el que emplean los propios reglamentos deportivos para definir un tiempo de vuelta aceptable.

Este último filtro condiciona sobre la propia variable respuesta y debe declararse como advierte la sección 3.2.4: todos los resultados de este capítulo se refieren a vueltas en aire limpio y bandera verde, no a la carrera completa, y la magnitud de los efectos estimados depende del umbral elegido.

El segundo filtro selecciona el par de *stints* de interés. La numeración de *stints* se desplaza cuando hay banderas rojas, salidas paradas o arranques en mojado, de modo que el primero y el segundo no siempre llevan los números uno y dos. Tomarlos literalmente habría descartado carreras completas, entre ellas Suzuka 2024 e Interlagos 2023. Se toma en su lugar el primer par de *stints* consecutivos con al menos cinco vueltas limpias cada uno.

El tercero es la exigencia de un mínimo de cinco vueltas limpias en cada *stint* del par. Nótese que no es redundante con el segundo aunque aparezca en su enunciado: el segundo filtro emplea ese mínimo como *criterio de selección* entre los pares de *stints* consecutivos disponibles, mientras que el tercero actúa como *condición de exclusión* cuando ningún par de la unidad lo cumple, en cuyo caso la unidad entera se descarta. La razón del mínimo es que, con menos de cinco vueltas, el ancla del primer *stint* se estimaría con muy poca precisión y la pendiente local del segundo no estaría definida.

El cuarto excluye las carreras en mojado, en las que el tiempo por vuelta responde a un régimen distinto y a una variable —el estado de la pista— que no se observa.

Tras los cuatro filtros, la tabla contiene 771 unidades y 17 093 observaciones. La partición temporal asigna a entrenamiento y validación las temporadas 2023 y 2024, con 10 031 observaciones, y a prueba la temporada 2025, con 7 062. Se trata de una partición estrictamente cronológica: ninguna observación de 2025 interviene en el ajuste, ni en la elección de hiperparámetros, ni en el cálculo de codificaciones, imputaciones o recortes.

4.3.2. Definición de la variable a predecir y generación de variables explicativas

4.3.2.1. Momento y horizonte de la predicción

La predicción se emite en el instante en que el coche completa su parada en boxes y reincorpora a pista con el nuevo juego montado. En ese instante se conocen: todas las vueltas del primer *stint* de ese piloto, la clasificación y los entrenamientos libres de ese fin de semana, las condiciones ambientales, el estado del pelotón y el compuesto que se acaba de montar.

Se consideran dos objetivos, que responden a preguntas distintas y se evalúan por separado.

4.3.2.2. Objetivo T_1 : tiempo por vuelta

Sea i la unidad piloto–carrera y $t = 1, \dots, H_i$ el índice de vuelta dentro de su segundo *stint*. El primer objetivo es la duración de esa vuelta, y_{it} , en segundos. Se predice para todo t desde el instante de la parada, sin observar ninguna vuelta intermedia. Se consideran dos variantes.

La variante *directa* predice y_{it} tal cual:

$$T_1^{\text{dir}} = y_{it}. \quad (4.1)$$

La variante *anclada* predice la desviación respecto al ritmo mediano observado en el primer *stint* de la misma unidad:

$$T_1^{\text{anc}} = y_{it} - \hat{a}_i, \quad \hat{a}_i = \text{mediana} \{y_{is} : s \in S_i^{(1)}\}, \quad (4.2)$$

donde $S_i^{(1)}$ es el conjunto de vueltas limpias del *primer stint* de la unidad i , es decir, del mismo piloto, en el mismo circuito y en el mismo año. Conviene subrayar los tres elementos, porque la definición debe quedar sin ambigüedad: la mediana se calcula con vueltas del propio piloto, de la propia carrera, y anteriores al instante de la predicción.

Tres justificaciones sostienen esta definición. Primera, $\hat{a}_i$ se construye íntegramente con información disponible en el instante de predecir, de modo que no introduce fuga temporal.

Segunda, se emplea la mediana y no la media, por robustez frente a las vueltas atípicas que sobreviven al filtro, ni la última vuelta observada, que incorporaría íntegramente la perturbación de un instante concreto en un estadístico que después condiciona toda la predicción.

Tercera, y decisiva, el anclaje afecta de manera muy distinta a las dos familias de modelos comparadas. El ritmo del primer *stint* guarda con el tiempo por vuelta del segundo una relación aproximadamente lineal de pendiente próxima a uno: si un coche rodaba a 95 segundos, rodará alrededor de 95 segundos más una corrección de pocas décimas. Un modelo lineal que disponga de $\hat{a}_i$ entre sus regresores estima esa pendiente y la extiende a circuitos cuyo ritmo no haya visto nunca; anclar el objetivo no le aporta nada, y la proposición 3.1 demuestra que sus predicciones no cambian en absoluto. Un modelo de árboles dispone del mismo regresor, pero por la proposición 3.3 solo puede utilizarlo para *partir* el espacio —«si el ritmo del primer *stint* supera 95 segundos, predecir tal constante»— y no para *sumarlo* a la predicción. Anclar el objetivo traslada esa suma a la definición de la variable a predecir y le evita tener que representarla. Se ajustan por ello ambas variantes: la comparación constituye un contraste empírico de la proposición.

4.3.2.3. Objetivo T_2 : incremento a dos vueltas

El segundo objetivo es el incremento del tiempo por vuelta entre la vuelta t y la vuelta $t + 2$ del segundo *stint*:

$$T_2 = \Delta_{it}^{(2)} = y_{i,t+2} - y_{i,t}, \quad (4.3)$$

predicho utilizando exclusivamente información disponible hasta la vuelta t inclusive. Se exige además continuidad de índice: la observación $t + 2$ debe pertenecer a la misma unidad y al mismo *stint*.

La justificación de este segundo objetivo es doble. Por un lado, responde a una pregunta operativa distinta: no cuánto tardará el coche, sino cómo va a evolucionar su ritmo a corto plazo, que es lo que informa la decisión de prolongar o acortar el *stint*. Por otro, y como observa la sección 3.2.5, la diferencia (4.3) elimina el nivel de unidad a_i de forma exacta y sin necesidad de estimarlo: es un objetivo autoanclado, de modo que la comparación entre T_1^{anc} y T_2 contrasta un anclaje estimado frente a uno exacto.

4.3.2.4. Generación de variables explicativas

Se construyen 54 variables explicativas agrupadas en nueve bloques, resumidos en la tabla 4.1. Los ocho primeros están disponibles en el instante de la parada y se emplean con ambos objetivos; el noveno recoge el estado observado dentro del segundo *stint* hasta la vuelta t y se emplea únicamente con T_2 . La definición precisa de cada variable, junto con su porcentaje de valores ausentes, se recoge en el anexo F.

Cinco construcciones merecen comentario por no ser lecturas directas de la fuente.

La primera es el propio ancla `s1_pace`, definida en (4.2), que figura además como variable explicativa en ambos objetivos.

La segunda es `s1_pace_rel`, que sitúa al piloto respecto al resto del campo:

$$\text{s1_pace_rel}_i = \hat{a}_i - \text{mediana} \{ \hat{a}_j : j \in \text{misma carrera} \}, \quad (4.4)$$

donde la mediana del segundo término se calcula *únicamente con primeros stints*. La precisión es esencial: calcularla con todas las vueltas de la carrera habría introducido en una variable explicativa información procedente de los segundos *stints*, que son precisamente lo que se quiere predecir.

CUADRO 4.1: Bloques de variables explicativas, con el número de variables de cada uno y el instante al que se refiere la información que contienen.

Bloque	Núm.	Contenido
Circuito	6	Características físicas del trazado y formato del fin de semana.
Clasificación	4	Resultado de la sesión de clasificación y posición de salida.
Entrenamientos libres	3	Ritmo y degradación observados en libres.
Primer <i>stint</i>	13	Ritmo, degradación, longitud, compuesto y tráfico sufrido durante el primer <i>stint</i> . Incluye el ancla $\hat{a}_i$.
Plan del segundo <i>stint</i>	6	Compuesto montado, edad del juego y vuelta de inicio: lo que el equipo decide en la parada.
Pelotón al parar	7	Configuración de huecos alrededor del coche en el instante de la parada.
Meteorología	5	Temperatura del aire y del asfalto, humedad, viento y presión.
Posición en el <i>stint</i>	3	Vuelta del segundo <i>stint</i> , edad del neumático y vuelta de carrera.
Estado hasta t (solo T_2)	8	Tiempo, desviación, pendiente local, último incremento e intervalo al precedente, observados hasta la vuelta t .

La tercera es la pendiente local de degradación dentro de un *stint*, estimada por mínimos cuadrados sobre las vueltas limpias observadas:

$$\hat{\delta} = \frac{\sum_s (a_s - \bar{a})(y_s - \bar{y})}{\sum_s (a_s - \bar{a})^2}, \quad (4.5)$$

con a_s la edad del neumático. Se calcula sobre el primer *stint* (**s1_deg**), sobre los entrenamientos libres (**fp_deg**) y, para T_2 , sobre las vueltas del segundo *stint* observadas hasta t (**pend_t**).

La cuarta es el bloque de pelotón, derivado de la diferencia de cada coche con el líder medida al inicio de cada vuelta. Ordenando los coches por esa diferencia se obtiene la posición virtual en ese instante, de la que se leen la densidad a dos y cinco segundos, el hueco libre por delante y la posición estimada tras reincorporarse, que suma a la diferencia con el líder la pérdida de boxes propia del circuito.

La quinta es el bloque de estado hasta t , exclusivo de T_2 : el tiempo observado en la vuelta t , su desviación respecto al ancla, la media de las tres últimas vueltas observadas, el incremento de la última vuelta $y_t - y_{t-1}$, la pendiente local (4.5) y el intervalo al coche precedente. Todas ellas cumplen la restricción de la sección 3.2.5: se calculan con vueltas de índice menor o igual que t .

Esa restricción se ha verificado mediante una comprobación automática que recorre las 54 variables y confirma que ninguna emplea información posterior al instante de predicción correspondiente a su objetivo.

4.3.3. Análisis exploratorio

4.3.3.1. Los tres objetivos

La figura 4.1 recoge la distribución de las tres definiciones y el anexo G recoge el descriptivo univariante completo de las 54 variables.

El objetivo T_1 directo tiene media 88,180 segundos y desviación típica 9,916, con recorrido de 69,196 a 117,541. Su distribución es claramente multimodal, con una moda por grupo de

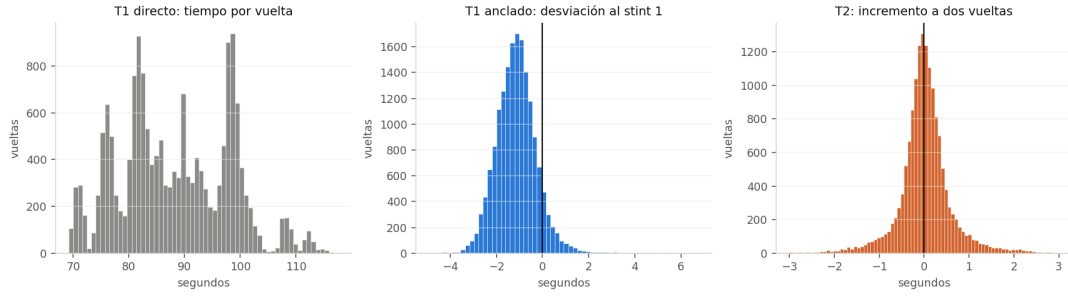


FIGURA 4.1: Distribución de los tres objetivos. A la izquierda T_1 directo, en el centro T_1 anclado y a la derecha T_2 . En los dos últimos se ha marcado el valor cero, que es la referencia respecto de la cual se define cada uno: el ritmo del primer *stint* en el caso de T_1 anclado y la ausencia de cambio de ritmo en el de T_2 .

circuitos: la desviación típica mide esencialmente diferencias entre trazados y no variación atribuible al neumático o al combustible. Su asimetría es 0,23 y su exceso de curtosis $-0,70$.

El objetivo T_1 anclado tiene media $-1,142$ segundos y desviación típica 0,884: el anclaje reduce la dispersión en un factor de 11. La media negativa es coherente con la física del problema, pues el segundo *stint* rueda con menos combustible y con goma nueva. Su asimetría es 0,75 y su exceso de curtosis 4,00: la cola derecha, la de las vueltas lentas, es más larga, lo que era esperable porque existe un límite físico a lo rápido que puede rodar un coche y ninguno a lo lento.

El objetivo T_2 tiene media 0,020 segundos y desviación típica 0,672. Su media prácticamente nula confirma que la diferencia elimina el nivel: a dos vueltas vista, el ritmo no cambia sistemáticamente. Lo destacable es su exceso de curtosis, 13,05, muy superior al de los otros dos objetivos: la distribución concentra masa en el centro y presenta colas muy pesadas, con valores extremos de $-7,221$ y $7,287$ segundos. Esta característica condiciona todo el tratamiento posterior y explica buena parte de los resultados de la sección 4.3.6.

4.3.3.2. Relación con las variables explicativas

Las figuras H.1 y H.2 representan las medias por tramo (3.27) con su intervalo al noventa y cinco por ciento.

Para T_1 anclado se observan cinco relaciones, todas ellas en la dirección esperada.

Como era de esperar, conforme avanza la vuelta de carrera el objetivo disminuye de forma prácticamente monótona, de $-0,807$ segundos en la vuelta 17 a $-1,544$ en la vuelta 50. Es el efecto del consumo de combustible, aproximadamente lineal, tal como corresponde a un gasto constante por vuelta.

La edad del neumático dibuja una curva en forma de uve muy tenue, con mínimo en torno a las seis vueltas ($-1,081$) y ascenso posterior hasta $-1,229$ en la vuelta treinta y uno. La amplitud total es de apenas 0,15 segundos, muy inferior a la del combustible. Esto es exactamente lo que cabe esperar de la superposición de dos efectos de signo contrario y magnitud comparable, y anticipa que ambos no serán separables sin más información.

Como era de esperar, cuantos más coches encadenados hay por delante en el instante de la parada, mayor es el tiempo por vuelta: el objetivo pasa de $-1,255$ segundos sin ningún coche a $-0,833$ con cinco o más, es decir, más de cuatro décimas de penalización. La densidad de coches a menos de dos segundos muestra el mismo patrón, de $-1,266$ a $-0,931$.

La longitud del primer *stint* presenta una relación decreciente marcada, de $-0,662$ con *stints* de ocho vueltas a $-1,735$ con *stints* de treinta y una. La interpretación es mecánica: cuanto más largo fue el primero, más tarde arranca el segundo y menos combustible lleva el coche, de modo que la variable actúa en buena medida como sustituto de la vuelta de carrera.

Por último, el compuesto montado ordena las medias como cabría esperar por el nivel —blando $-1,424$, duro $-1,116$ — pero esa ordenación mide sobre todo *cuándo* se monta cada compuesto y no su diferencia de agarre, puesto que el blando se monta en *stints* cortos y tardíos.

Para T_2 el patrón dominante es distinto y mucho más nítido: es reversión a la media. Las tres variables más informativas son las que miden el estado actual respecto a su propio nivel.

La desviación actual respecto al ancla presenta una relación monótona decreciente y de gran amplitud: cuando el coche rueda 2,48 segundos por debajo de su ancla, el incremento medio a dos vueltas es $+0,229$ segundos; cuando rueda por encima ($+0,18$), es $-0,593$. Es decir, como era de esperar, tras una vuelta anormalmente rápida el coche tiende a ralentizarse y tras una anormalmente lenta tiende a recuperar.

El incremento de la última vuelta muestra el mismo comportamiento: de $+0,131$ segundos cuando la vuelta anterior fue 0,84 segundos más rápida, a $-0,637$ cuando fue 0,83 segundos más lenta. Y la pendiente local hasta t va de $+0,096$ a $-0,196$ conforme la pendiente pasa de negativa a positiva.

Las variables conocidas en el instante de la parada, en cambio, apenas se relacionan con T_2 : la vuelta de carrera oscila entre $-0,018$ y $+0,068$ sin patrón claro, y la edad del neumático entre $-0,014$ y $+0,084$. Esto anticipa un resultado importante: a dos vueltas vista, lo que informa es el estado inmediato del coche y no las condiciones de partida.

4.3.4. Tratamiento de variables

4.3.4.1. Valores ausentes

Los ausentes se concentran en tres bloques con orígenes distintos. Las variables meteorológicas faltan en el 10,37 por ciento de las observaciones, porque la estación no publicó registro en esa ventana. Las de pelotón faltan en el 9,60 por ciento, por cobertura incompleta del recurso de intervalos. Las de entrenamientos libres faltan en el 4,66 por ciento, porque el piloto no completó vueltas útiles, lo que ocurre sobre todo en fines de semana al esprint, donde solo hay una sesión libre.

Este último caso es informativo en el sentido de la sección 3.2.7: la ausencia correlaciona con el formato del fin de semana. Se adopta, por ello, un tratamiento diferenciado. Los modelos de árboles gestionan la ausencia de forma nativa, enviando la observación al lado del corte que minimiza la pérdida. Para los modelos lineales generalizados y que no admiten ausencias, se imputa por la mediana del tramo de ajuste y se añade una variable indicadora de ausencia, de manera que el modelo pueda estimar por separado el efecto de faltar. Ambas cantidades —mediana y estructura de indicadoras— se calculan exclusivamente con las temporadas 2023 y 2024. Conviene precisar el alcance de esa afirmación: dentro del tramo de ajuste la imputación es global, de modo que un valor ausente de 2023 se imputa con una mediana que incorpora observaciones de 2024. Se trata de una fuga de información *dentro* del ajuste, no hacia la prueba, y por tanto no invalida la estimación del error de generalización, que es lo que se reporta; su único efecto sería un error de ajuste ligeramente optimista. La muestra de prueba, en cambio, se imputa con cantidades calculadas únicamente con 2023 y 2024, sin que ninguna observación de 2025 intervenga en su propia imputación.

4.3.4.2. Valores atípicos

La figura 4.2 recoge las variables con mayor proporción de valores marcados por las dos reglas de detección; el anexo C reproduce los porcentajes exactos.

Su lectura exige distinguir tres situaciones que las reglas confunden.

La primera es la de las variables discretas ordinales de pocos niveles. La variable `s1_compound_ord` encabeza la tabla con un 31,71 por ciento de atípicos según (3.28), cifra que

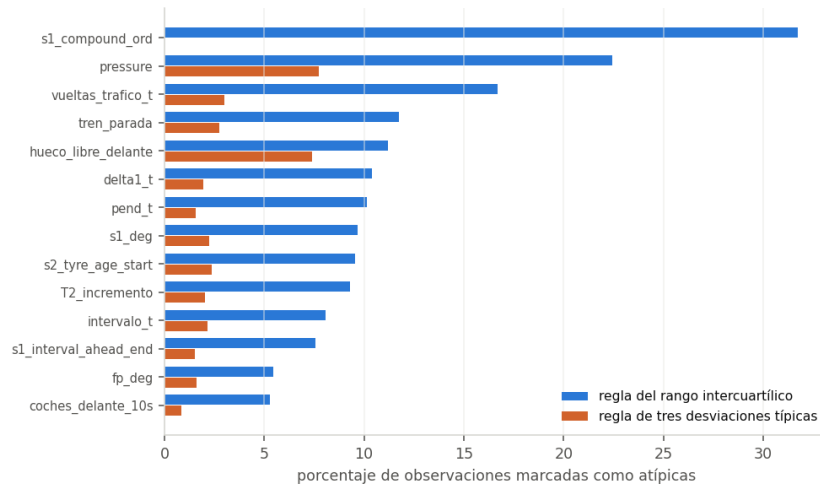


FIGURA 4.2: Porcentaje de observaciones marcadas como atípicas por cada regla. La discrepancia entre ambas es indicio de asimetría, no de error de medida.

carece de contenido: la variable toma tres valores y sus cuartiles primero y tercero coinciden, de modo que el rango intercuartílico es nulo y la regla marca todo valor distinto de la mediana. La regla de las tres desviaciones típicas da, coherentemente, cero por ciento.

La segunda es la de las variables asimétricas por construcción. *vueltas_trafico_t*, con 16,69 por ciento y asimetría 3,64, y *tren_parada*, con 11,73 por ciento y asimetría 3,20, cuentan con valores con poca probabilidad de ocurrencia: su valor modal es cero y los valores altos corresponden a situaciones reales de tráfico. Eliminarlos suprimiría precisamente el fenómeno que se quiere estudiar.

La tercera es la de los valores extremos legítimos. La presión atmosférica presenta un 22,43 por ciento de atípicos por (3.28) y un 7,73 por ciento por la regla de tres desviaciones, con mínimo de 779,2 hectopascales y asimetría $-2,32$. Ese mínimo corresponde a Ciudad de México, a más de dos mil doscientos metros de altitud, donde la presión real es esa. Un procedimiento automático de eliminación habría descartado esa carrera completa.

4.3.4.3. Comparación de tratamientos

En lugar de decidir a priori, se ajustan todos los modelos bajo los cuatro tratamientos de la sección 3.2.7: ninguno, recorte a los percentiles 1 y 99, eliminación por (3.28) sobre el objetivo y transformación logarítmica de las diez variables de mayor asimetría. Los parámetros de cada transformación se calculan siempre sobre 2023 y 2024. Los resultados completos de esas pruebas se recogen en el anexo A; aquí se resume únicamente la conclusión que determina la decisión adoptada.

El resultado es nítido y confirma la predicción teórica de la sección 3.2.7. El tratamiento de atípicos beneficia sustancialmente a los modelos que minimizan el error cuadrático y apenas afecta a los que minimizan el error absoluto. Para el objetivo T_2 , el *boosting* con pérdida cuadrática mejora un 13,6 por ciento al eliminar atípicos, y el GLM gaussiano un 9,3 por ciento; el mismo *boosting* con pérdida absoluta mejora un 0,2 por ciento. Para T_1 anclado las cifras son 7,6, 3,8 y 0,0 por ciento respectivamente.

La explicación es la ya dada: bajo pérdida cuadrática la influencia de una observación crece linealmente con su error, mientras que bajo pérdida absoluta está acotada por la derivada $\text{sign}(y - f)$. Eliminar atípicos y emplear una pérdida robusta resuelven el mismo problema, y hacer ambas cosas es redundante. Se adopta, en consecuencia, la segunda: no se elimina ninguna observación, y la robustez se obtiene mediante la función de pérdida. Esta decisión

tiene además la ventaja de no condicionar sobre la variable respuesta, que es lo que hace la eliminación y lo que la sección 3.2.4 señala como problemático.

4.3.4.4. Transformaciones

No se aplica ninguna transformación monótona a los regresores de los modelos de árboles, porque el criterio (3.15) depende únicamente del orden de los valores y cualquier transformación creciente produce el mismo árbol. Los resultados del anexo A lo confirman: el efecto del tratamiento logarítmico sobre el *boosting* es de milésimas.

Para los modelos lineales generalizados se tipifican todas las variables, condición necesaria para que los coeficientes de (3.4) sean comparables entre regresores. Se aplica además, previamente a la tipificación, un filtro de rango mediante factorización QR con pivoteo por columnas, que descarta las columnas linealmente dependientes. Este paso resultó necesario: sin él, la eliminación de filas atípicas generaba colinealidad exacta entre indicadoras de ausencia y producía coeficientes divergentes.

4.3.5. Ajuste de modelos analíticos

Se ajustan a cada uno de los tres objetivos tres predictores de referencia y cuatro modelos, todos ellos con el mismo conjunto de variables explicativas y la misma partición temporal.

Predictor constante. Se emplea la mediana del tramo de ajuste, que es el predictor constante óptimo bajo la métrica principal. Establece el suelo absoluto: un modelo que no lo supere no ha extraído información alguna de los regresores.

Constante por piloto y circuito. Refinamiento del anterior, que estima una mediana distinta para cada par piloto-circuito observado durante el ajuste. Cuando un par no aparece en el tramo de ajuste, hereda la mediana de su circuito y, en última instancia, la global. Es una referencia mucho más exigente, pues incorpora las dos variables categóricas más informativas del problema.

Persistencia. Predice que la magnitud de interés no cambiará respecto de su último valor observado. Para el objetivo T_1 equivale a predecir que el segundo *stint* rodará al ritmo del primero, es decir, a predecir cero en la escala anclada; para el objetivo T_2 , a predecir un incremento nulo, esto es, que la vuelta $t+2$ durará lo mismo que la vuelta t . Sobre T_2 se añade una segunda variante, que repite el último incremento observado del mismo alcance, $y_t - y_{t-2}$.

GLM gaussiano con enlace identidad. Estimado por (3.9), que en este caso converge en una iteración y coincide con (3.4).

GLM Gamma con enlace logarítmico. Requiere soporte estrictamente positivo. El objetivo T_1 directo lo cumple, por ser un tiempo. Los objetivos T_1 anclado y T_2 no: toman valores negativos por construcción, con mínimos de $-4,914$ y $-7,221$ segundos respectivamente. El modelo se declara por tanto no aplicable a esos dos objetivos, y así consta en la tabla I.1. Cabría forzar la positividad desplazando el objetivo mediante una constante, pero se ha descartado esa vía porque el valor del desplazamiento es arbitrario y las estimaciones de los coeficientes dependen de él: con enlace logarítmico, sumar una constante a la respuesta altera la interpretación multiplicativa del modelo y no existe criterio objetivo para fijarla.

Boosting con pérdida cuadrática y con pérdida absoluta. Ambas variantes emplean la implementación LightGBM descrita en la sección 3.1.2.4, con la misma configuración de partida: $M = 700$ árboles, $\eta = 0,03$, $J = 31$ hojas, tamaño mínimo de hoja 40, submuestreo de filas y de columnas 0,8 y penalización $\lambda = 5$. Son valores de referencia habituales, no elegidos con los datos, de modo que en estas dos variantes no interviene ninguna muestra de validación: los parámetros se estiman sobre las temporadas 2023 y 2024 completas y el error se mide sobre 2025. Se ajustan las dos pérdidas para contrastar el argumento de la sección 3.2.9.

Boosting con hiperparámetros optimizados. Es la única variante que requiere el esquema de tres muestras de la sección 3.2.8. Sobre la configuración de pérdida absoluta se realiza una búsqueda bayesiana de treinta ensayos con el estimador de Parzen (3.29), en la que los parámetros se estiman con la temporada 2023 y las configuraciones se comparan según el error obtenido sobre la temporada 2024, que actúa por tanto como muestra de validación. Elegida la mejor configuración, el modelo se reajusta sobre 2023 y 2024 conjuntamente y se evalúa sobre 2025. La temporada de prueba no interviene en ningún momento de la elección. El espacio explorado y las configuraciones ganadoras se recogen en el anexo B.

4.3.6. Validación de resultados

La tabla I.1 recoge las tres métricas de (3.30) sobre el conjunto de ajuste y sobre el de prueba, para los tres predictores de referencia, los cuatro modelos ajustados y los tres objetivos.

4.3.6.1. Cuál es el mejor objetivo

La comparación entre las tres definiciones del objetivo es la conclusión principal de este capítulo, y conviene enunciarla con precisión porque las métricas no son directamente comparables entre objetivos: T_1 y T_2 miden cosas distintas y su escala es distinta. La comparación legítima es la de cada modelo contra el predictor constante *del mismo objetivo*.

Para T_1 directo, el predictor constante comete 7,999 segundos y el mejor modelo 0,583: una reducción del 92,7 por ciento. Nótese, de paso, que su R^2 de prueba es negativo ($-0,035$): como se explicó en la sección 3.2.9, la constante se calcula sobre el tramo de ajuste y no coincide con la media de la muestra de prueba, de modo que el predictor constante puede quedar por debajo de esa media y arrojar un coeficiente negativo. Ahora bien, esa cifra es engañosa por el motivo que advierte la sección 3.2.9: la variación del tiempo por vuelta entre circuitos es de unos 48 segundos, de modo que basta con identificar el circuito para obtener un R^2 de 0,99 sin haber explicado nada de lo que ocurre dentro de un *stint*. Y en efecto, todos los modelos salvo los dos peores superan $R^2 = 0,97$ sobre este objetivo.

Para T_1 anclado, el predictor constante comete 0,690 segundos y el mejor modelo 0,525: una reducción del 23,9 por ciento, con $R^2 = 0,310$. Esta cifra sí es informativa, porque el objetivo ya no contiene la variación entre circuitos.

Para T_2 , el predictor constante comete 0,394 segundos y el mejor modelo 0,351: una reducción del 10,8 por ciento, con $R^2 = 0,251$. El incremento a dos vueltas es, por tanto, sustancialmente más difícil de predecir que el nivel anclado, lo cual es coherente con lo observado en el análisis bivalente: la señal disponible es esencialmente reversión a la media.

4.3.6.2. Comparación con los predictores de referencia

Antes de comparar modelos entre sí conviene situarlos frente a los predictores sencillos, que es la pregunta que un lector escéptico plantea primero: ¿cuánto de lo que consigue el modelo procede de la información que extrae y cuánto de la mera inercia del fenómeno?

La respuesta depende del objetivo y arroja dos resultados negativos que merecen declararse. El primero es que **la constante por piloto y circuito no mejora a la constante global**

en ninguno de los tres objetivos: sobre T_1 anclado obtiene 0,745 frente a 0,690, y sobre T_2 , 0,406 frente a 0,394. El motivo es que solo el 57,9 por ciento de los pares piloto–circuito de la temporada de prueba había sido observado durante el ajuste —la parrilla cambia de un año a otro y el calendario incorpora circuitos nuevos—, y los pares que sí aparecen lo hacen con una mediana de 24 observaciones, insuficientes para estimar con precisión. Es un ejemplo directo del término de varianza de (3.16): al estimar una constante por grupo se elimina sesgo pero se paga en inestabilidad, y aquí el pago excede a la ganancia.

El segundo es más relevante para el objetivo T_2 . **El predictor de persistencia —suponer que la vuelta $t + 2$ durará lo mismo que la vuelta t — comete 0,393 segundos, y supera a los dos modelos que estiman la media condicional:** el GLM gaussiano (0,438) y el *boosting* con pérdida cuadrática (0,444). Solo los modelos ajustados con pérdida absoluta lo baten, y el mejor de ellos, el *boosting* optimizado, lo hace con 0,351 segundos, una mejora del 10,7 por ciento. Esa cifra, y no la comparación contra la constante global, es la medida honesta de lo que aporta el modelo sobre T_2 .

Conviene añadir un tercer resultado, esta vez con signo positivo. Repetir el último incremento observado —predecir Δ mediante $y_t - y_{t-2}$ — comete 0,623 segundos, muy por encima de predecir un incremento nulo. La diferencia entre ambas referencias confirma que el proceso no tiene inercia sino *reversión*: tras una vuelta anormalmente rápida lo esperable no es otra igual de rápida, sino una corrección en sentido contrario. Es exactamente el patrón que anticipaba el análisis bivalente de la sección 4.3.3 y que el modelo ganador reproduce.

Sobre T_1 la lectura es distinta. La persistencia —predecir que el segundo *stint* rodará al ritmo del primero— comete 1,341 segundos, muy por encima de cualquier modelo, porque ignora las dos fuerzas descritas en la sección 2.3.1: el coche va a rodar sistemáticamente más rápido por llevar menos combustible y goma nueva. El margen que los modelos obtienen sobre esta referencia, de 1,341 a 0,525 segundos, sí es por tanto una medida de información genuinamente aprendida.

4.3.6.3. Contraste de la proposición sobre el anclaje

El resultado que valida la teoría del capítulo 3 es el siguiente. El GLM gaussiano obtiene exactamente 0,5829 segundos de error de prueba sobre T_1 directo y exactamente 0,5829 sobre T_1 anclado, con idénticos valores de MAE y RMSE hasta la cuarta cifra decimal.

Esta coincidencia no es casual: es la proposición 3.1 verificándose numéricamente. Como el ancla `s1_pace` figura entre los regresores, anclar el objetivo es una transformación que el estimador de mínimos cuadrados absorbe exactamente, desplazando en una unidad el coeficiente correspondiente y dejando los residuos invariantes.

Para el *boosting*, en cambio, el anclaje es decisivo. Con pérdida cuadrática, el error pasa de 0,915 sobre T_1 directo a 0,600 sobre T_1 anclado, una reducción del 34,4 por ciento; con pérdida absoluta, de 0,867 a 0,525, una reducción del 39,5 por ciento. La causa es la proposición sobre extrapolación: un modelo de la forma (3.14) no puede predecir fuera del recorrido observado del objetivo, y el recorrido de los tiempos por vuelta de 2025 no coincide con el de 2023 y 2024.

De aquí se sigue una conclusión que a primera vista resulta contraintuitiva: sobre el objetivo directo el modelo lineal bate con holgura al *boosting* (0,583 frente a 0,867), y solo tras anclar se invierte la ordenación (0,583 frente a 0,525). No es que un modelo sea mejor que otro en abstracto: es que la definición del objetivo determina qué clase de modelos puede resolverlo.

4.3.6.4. Efecto de la función de pérdida

El segundo contraste teórico se refiere a la alineación entre la pérdida de ajuste y la métrica de evaluación. Sobre T_1 anclado, el *boosting* pasa de 0,600 con pérdida cuadrática a 0,525 con

pérdida absoluta, una mejora del 12,5 por ciento; sobre T_2 , de 0,444 a 0,368, un 16,9 por ciento. En ambos casos la mejora se obtiene sin cambiar ni una variable ni un hiperparámetro: únicamente el funcional que se estima.

El caso de T_2 es especialmente ilustrativo. El GLM gaussiano obtiene 0,438 de MAE, *peor* que el predictor constante (0,394), y sin embargo obtiene 0,603 de RMSE frente a 0,631 del constante, y $R^2 = 0,087$ frente a $-0,000$. Es decir: el modelo sí extrae señal, pero como estima la media condicional y se le evalúa con una métrica que premia la mediana, en una distribución con exceso de curtosis 13,05 resulta penalizado. Este es exactamente el fenómeno anticipado tras (3.1), y muestra por qué reportar una sola métrica puede conducir a descartar un modelo que sí funciona.

4.3.6.5. Sobreajuste

La última columna de la tabla I.1 recoge el cociente entre el error de prueba y el de ajuste.

El caso más ilustrativo es el de la constante por piloto y circuito. Su error de ajuste sobre T_1 anclado es 0,441 segundos, mejor que el de cualquier modelo lineal, y sin embargo su error de prueba, 0,745, es el peor de la tabla: el cociente asciende a 1,69. Al estimar una constante para cada uno de los 384 pares observados se está ajustando un modelo con 384 parámetros sobre 10031 observaciones, lo que reproduce bien el pasado y generaliza mal. Es la ilustración más directa de que un error de ajuste bajo no informa por sí solo sobre la capacidad de generalización.

Los modelos de *boosting* presentan cocientes entre 1,8 y 3,2: con la configuración de partida reproducen los datos de ajuste con un error entre 0,25 y 0,29 segundos que no trasladan a la prueba. Los modelos lineales, en el otro extremo, tienen cocientes en torno a 1,1, lo que es coherente con su número de parámetros y su forma funcional impuesta.

La búsqueda de hiperparámetros actúa sobre esa brecha. En T_2 , el *boosting* optimizado pasa de un error de ajuste de 0,274 a 0,341 —es decir, ajusta *peor* los datos que ya ha visto— y simultáneamente mejora el error de prueba de 0,368 a 0,351. Es la manifestación directa del compromiso entre sesgo y varianza: la regularización sacrifica ajuste dentro de muestra para ganar fuera de ella.

4.3.6.6. Modelo ganador

El modelo ganador es el *boosting* con pérdida absoluta, en su configuración de partida para el objetivo T_1 anclado (0,525 segundos, frente a 0,532 de la versión optimizada) y en su versión con hiperparámetros optimizados para el objetivo T_2 (0,351 segundos, frente a 0,368). Que la optimización de hiperparámetros mejore en un objetivo y no en el otro no es contradictorio: el espacio de búsqueda se recorre con treinta ensayos y la configuración elegida se valida sobre una única temporada, de modo que la diferencia de siete milésimas sobre T_1 anclado está dentro del ruido de esa validación.

En ambos objetivos el *boosting* supera al GLM gaussiano, que es el mejor de los modelos lineales (0,583 y 0,438 respectivamente).

La diferencia entre ambas familias es, sin embargo, moderada: el *boosting* mejora al GLM en un 8,8 por ciento sobre T_1 anclado. Dado que el GLM ofrece interpretación directa por parámetros y un sobreajuste mucho menor, esa diferencia es el precio que habría que pagar por prescindir de la interpretabilidad, y no una superioridad cualitativa.

4.3.7. Interpretación de los resultados del modelo ganador

Se interpretan los resultados por las dos vías de la sección 3.2.9: los parámetros del GLM y los valores de Shapley del *boosting*. El objetivo es contrastar ambas con las intuiciones que el análisis bivalente de la sección 4.3.3 había anticipado.

4.3.7.1. Objetivo T_1 anclado

La figura 4.3 representa la importancia global según (3.23); el anexo D recoge los valores exactos.

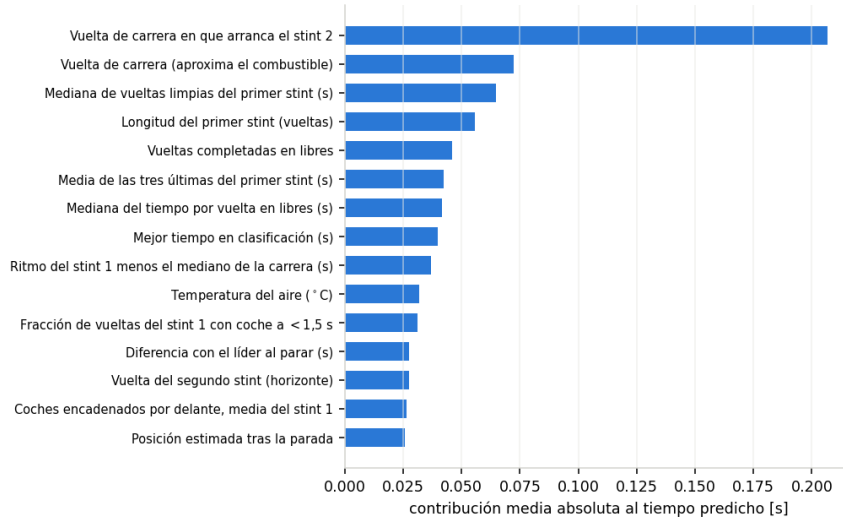


FIGURA 4.3: Importancia global según los valores de Shapley para el objetivo T_1 anclado, medida como contribución media absoluta en segundos.

La variable dominante es `s2_start_lap`, la vuelta de carrera en que arranca el segundo *stint*, con una contribución media absoluta de 0,207 segundos, tres veces superior a la siguiente. Le siguen `lap_number` (0,072) y `s1_pace` (0,065).

Esta atribución valida la intuición del análisis bivalente. Las dos primeras variables miden esencialmente lo mismo —cuánto combustible lleva el coche— y eran precisamente las que mostraban la relación más limpia y de mayor amplitud con el objetivo, con un recorrido de 0,74 segundos entre el primer y el último decil. Que el modelo, sin que se le indique nada, concentre su capacidad explicativa en el consumo de combustible es el contraste más directo posible entre lo esperado y lo aprendido.

La tabla 4.2 recoge los coeficientes del GLM gaussiano, tipificados. Su lectura confirma el mismo diagnóstico por otra vía y añade información que el SHAP no proporciona: el signo y su significación.

Los dos coeficientes de mayor magnitud son `s1_last3` (+2,296) y `s1_pace` (−2,182), con signos opuestos y magnitudes casi idénticas. Esta configuración es el síntoma característico de la colinealidad descrita en (3.5): ambas variables miden el ritmo del primer *stint* y su correlación es muy alta, de modo que el modelo estima esencialmente su *diferencia*, que es una medida de degradación al final del primer *stint*. Los coeficientes individuales no admiten, por tanto, lectura separada; su combinación sí.

Entre los coeficientes interpretables aisladamente, `s1_pace_rel` (−0,370, con $z = -21,0$) indica que un piloto una desviación típica más lento que la mediana de su carrera tiene un objetivo 0,370 segundos menor, es decir, mejora más respecto a su propio primer *stint*; `s2_start_lap` (−0,340, $z = -15,2$) recoge el efecto del combustible ya identificado; y `abrasion` (−0,363, $z = -17,2$) indica que en circuitos más abrasivos la mejora del segundo *stint* es mayor, lo que es coherente con que allí la goma nueva marque más diferencia.

CUADRO 4.2: Diez coeficientes de mayor magnitud del GLM gaussiano para el objetivo T_1 anclado. Las variables están tipificadas, de modo que $\hat{\beta}_j$ es el cambio esperado en el objetivo, en segundos, ante un aumento de una desviación típica en el regresor, manteniendo los demás constantes.

Variable	Descripción	$\hat{\beta}_j$	$\hat{e}e$	z
s1_last3	Media de las tres últimas del primer stint (s)	2,2958	0,2343	9,8
s1_pace	Mediana de vueltas limpias del primer stint (s)	-2,1815	0,2591	-8,4
const		-1,0600	0,0073	-145,0
s1_pace_rel	Ritmo del stint 1 menos el mediano de la carrera (s)	-0,3701	0,0176	-21,0
abrasion	Abrasividad del asfalto (escala 1–5)	-0,3627	0,0211	-17,2
s2_start_lap	Vuelta de carrera en que arranca el stint 2	-0,3403	0,0225	-15,2
qual_best	Mejor tiempo en clasificación (s)	-0,2742	0,1035	-2,6
aero_load	Carga aerodinámica del trazado (1–5)	0,1910	0,0351	5,4
s1_len	Longitud del primer stint (vueltas)	-0,1849	0,0382	-4,8
pressure	Presión atmosférica (hPa)	0,1799	0,0221	8,1

4.3.7.2. Objetivo T_2

La figura 4.4 recoge la atribución para el objetivo de incremento, cuyos valores exactos figuran igualmente en el anexo D.

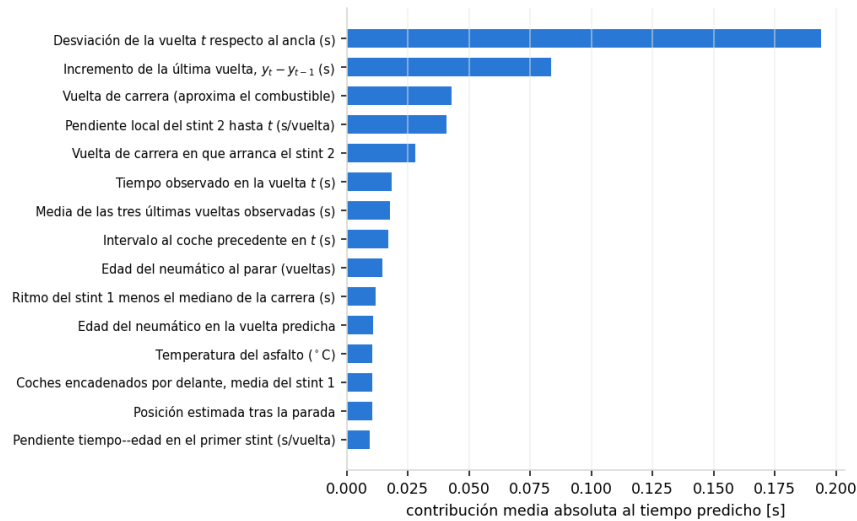


FIGURA 4.4: Importancia global según los valores de Shapley para el objetivo T_2 .

Aquí la jerarquía es completamente distinta y confirma punto por punto lo anticipado por el bivalente. Las cuatro variables más importantes son `dev_t` (0,194), `delta1_t` (0,084), `lap_number` (0,043) y `pend_t` (0,041). Tres de las cuatro pertenecen al bloque de estado observado hasta t , es decir, describen cómo viene rodando el coche en ese momento, y no las condiciones de partida.

La figura E.1 representa la contribución $x_{ij} \mapsto \varphi_j(x_i)$ para las variables principales, y su forma es inequívocamente decreciente: cuanto más rápido rueda el coche respecto a su ancla, mayor es la contribución negativa al incremento, esto es, mayor la tendencia a ralentizarse.

El modelo ha aprendido la reversión a la media sin que se le haya impuesto ninguna forma funcional.

Los coeficientes del GLM, recogidos en el anexo D, muestran de nuevo el patrón de colinealidad: `y_t` ($-4,789$) y `media3_t` ($+3,907$) tienen signos opuestos y magnitudes comparables, de modo que el modelo estima su diferencia, que es una medida de cuánto se desvía la última vuelta de la tendencia reciente. El coeficiente de `dev_t` ($-0,245$, $z = -12,6$), que sí es interpretable aisladamente, confirma la reversión: una desviación típica por encima del ancla implica un incremento 0,245 segundos menor.

4.3.7.3. Conclusión de la interpretación

Las dos vías de interpretación coinciden entre sí y con el análisis exploratorio previo, lo que constituye la validación de la capacidad explicativa. Para el objetivo de nivel, el modelo se apoya en el consumo de combustible y en el ritmo propio del piloto; para el objetivo de incremento, en el estado inmediato del coche respecto a su propia referencia. En ninguno de los dos casos la importancia de una variable resulta inexplicable a la luz de las medias por tramo, que es la comprobación que exige la sección 3.2.9.

Debe recordarse, no obstante, la limitación señalada en la sección 3.1.2.6: los valores de Shapley explican el modelo y no el fenómeno. La configuración de coeficientes opuestos que se observa en ambos GLM lo ilustra: bajo colinealidad, el reparto entre variables depende de cuál haya utilizado el algoritmo, y solo su combinación admite lectura sustantiva.

4.4. Limitaciones

El filtro del ciento diez por ciento condiciona sobre la variable respuesta, de modo que todos los resultados se refieren a vueltas en aire limpio y bandera verde, y la magnitud de los efectos depende del umbral elegido, cuya sensibilidad no se ha explorado.

Las variables de pelotón se derivan de la diferencia con el líder medida al inicio de la vuelta. No captan por tanto los adelantamientos producidos dentro de la vuelta ni detectan a los coches doblados, cuya diferencia con el líder es muy distinta de la de los coches con los que comparten pista.

La partición temporal es única: se ajusta con 2023 y 2024 y se evalúa con 2025. El error de prueba se estima, por tanto, sobre una sola temporada, y su variabilidad no se ha cuantificado. Una validación con ventana expansiva sobre más temporadas reduciría esa incertidumbre a costa de multiplicar el cómputo.

Por último, la separación entre degradación del neumático y consumo de combustible no es identificable con estos datos: dentro de un *stint* la edad del neumático y la vuelta de carrera avanzan a la vez, de modo que tras eliminar el efecto de unidad ambas variables resultan exactamente colineales. Lo observado en el análisis bivalente —una curva en uve de amplitud reducida— es compatible con esa indeterminación y no la resuelve.

Capítulo 5

Conclusiones

En este trabajo se ha abordado el problema de predecir el tiempo por vuelta de un piloto durante el segundo *stint* de un Gran Premio de Fórmula 1, emitiendo la predicción en el instante de la primera parada en boxes y empleando exclusivamente la información disponible en ese momento. A partir de datos públicos de cronometraje de las temporadas 2023 a 2025 se ha construido una tabla de 17 093 vueltas agrupadas en 771 unidades piloto–carrera, se han definido tres variables de respuesta alternativas y se han comparado dos familias de modelos, ajustadas con las temporadas 2023 y 2024 y evaluadas sobre la de 2025.

5.1. Conclusiones

Sobre la definición de la variable a predecir. La conclusión principal es que la definición del objetivo condiciona qué familia de modelos puede resolverlo, hasta el punto de invertir la ordenación entre ellas. Sobre el tiempo por vuelta en su escala original, el modelo lineal generalizado gana con holgura al *boosting* (0,583 frente a 0,867 segundos de error absoluto medio), porque puede estimar un coeficiente próximo a uno sobre el ritmo observado en el primer *stint* y extenderlo fuera del recorrido conocido, mientras que un modelo de árboles solo puede emplear ese regresor para partir el espacio. Al anclar el objetivo —restarle ese ritmo de referencia— la ordenación se invierte y el *boosting* pasa a ganar (0,525 frente a 0,583). El modelo lineal, en cambio, obtiene exactamente el mismo error en ambos casos, tal como establece la proposición 3.1, lo que constituye una verificación empírica de un resultado teórico.

Sobre la mejor estrategia de modelización. Una vez definido adecuadamente el objetivo, la mejor estrategia para abordar el problema consiste en emplear modelos de tipo *gradient boosting* ajustados con pérdida absoluta y con hiperparámetros optimizados mediante búsqueda bayesiana. La elección de la función de pérdida resulta tan influyente como la del algoritmo: cambiar la pérdida cuadrática por la absoluta reduce el error un 12,5 por ciento sobre el objetivo anclado y un 16,9 por ciento sobre el incremento, sin modificar ni una variable ni un hiperparámetro. La razón es que la métrica de evaluación empleada, el error absoluto medio, la minimiza la mediana condicional y no la media.

Sobre el tratamiento de valores atípicos. El tratamiento de atípicos beneficia sustancialmente a los modelos que minimizan el error cuadrático —hasta un 13,6 por ciento— y resulta prácticamente irrelevante para los que minimizan el error absoluto, cuya mejora no llega al 0,3 por ciento. Eliminar observaciones extremas y emplear una pérdida robusta son, por tanto, dos soluciones al mismo problema, y basta con adoptar una de ellas. Se ha optado por la segunda, que además evita condicionar sobre la variable de respuesta.

Sobre la comparación con predictores sencillos. Los modelos ajustados superan a los predictores de referencia, pero por márgenes que conviene declarar con precisión. Sobre el incremento a dos vueltas, el predictor de persistencia —suponer que el ritmo no cambiará— comete 0,393 segundos y el mejor modelo 0,351, una mejora del 10,7 por ciento. El predictor constante calculado por piloto y circuito no mejora al constante global en ninguno de los tres objetivos, lo que se explica porque solo el 57,9 por ciento de los pares piloto–circuito de la temporada de prueba había sido observado durante el ajuste. Estos resultados negativos son parte del aprendizaje del estudio: buena parte de la señal disponible a corto plazo es inercia del propio fenómeno, y cuantificar cuánta es condición necesaria para valorar la aportación del modelo.

Sobre la interpretación. Las dos vías de interpretación empleadas —los coeficientes del modelo lineal y los valores de Shapley del *boosting*— coinciden entre sí y con lo anticipado por el análisis exploratorio. Para el objetivo de nivel, el modelo se apoya en las variables que recogen el consumo de combustible; para el objetivo de incremento, en el estado inmediato del coche respecto de su propia referencia, reproduciendo un patrón de reversión a la media que no se le había impuesto.

5.2. Líneas futuras

Como futuros trabajos se proponen cuatro líneas.

La primera es ampliar la validación temporal. La partición empleada es única —ajuste con 2023 y 2024, prueba con 2025—, de modo que el error de generalización se estima sobre una sola temporada. Una validación con ventana expansiva sobre más temporadas permitiría cuantificar la variabilidad de esa estimación.

La segunda es incorporar el tratamiento nativo de variables categóricas. El problema contiene dos variables categóricas muy relevantes, el circuito y el piloto, que aquí se han codificado numéricamente mediante la contracción (3.25). Implementaciones de *boosting* con tratamiento específico de categóricas, como CatBoost [30], podrían explotarlas mejor.

La tercera es enriquecer la información sobre el tráfico. Las variables de pelotón se derivan de la diferencia con el líder medida al inicio de cada vuelta, lo que no capta los adelantamientos producidos dentro de la vuelta ni distingue a los coches doblados. Disponer de la posición en pista instante a instante permitiría estimar con mayor precisión el coste del aire sucio.

La cuarta es extender el sistema de la predicción a la decisión. Predecir la curva de tiempos por vuelta es el ingrediente principal de la condición (2.1) que gobierna el *undercut*, pero convertir esa predicción en una recomendación de estrategia exige además modelar la posición en pista, los adelantamientos y la aparición de coches de seguridad, e incorporar la incertidumbre de la propia predicción, en la línea de los simuladores estocásticos de [19].

Apéndice A

Comparación de tratamientos de valores atípicos

Se recogen aquí los resultados completos del experimento descrito en la sección 4.3.4, en el que cada modelo se ajusta bajo los cuatro tratamientos de valores atípicos contemplados en la sección 3.2.7: no aplicar ninguno, recortar a los percentiles 1 y 99, eliminar las observaciones que la regla (3.28) marca sobre la variable de respuesta, y aplicar la transformación logarítmica a las diez variables de mayor asimetría. Los parámetros de cada transformación se calculan siempre sobre las temporadas 2023 y 2024, y en ningún caso se altera la muestra de prueba.

CUADRO A.1: Error absoluto medio sobre 2025 según el tratamiento de atípicos aplicado. La última columna es la mejora relativa del mejor tratamiento sobre no aplicar ninguno.

Modelo	Ninguno	Recorte	Eliminación	Logaritmo	Mejora (%)
<i>T₁ directo</i>					
Predictor constante (referencia)	7,9986	7,9986	7,9986	7,9986	0,0
GLM Gamma, enlace logarítmico	0,7444	0,7433	0,7444	0,7671	0,2
GLM gaussiano, enlace identidad	0,5829	0,5824	0,5829	0,5842	0,1
<i>Boosting</i> , pérdida cuadrática	0,9146	0,8421	0,9146	0,9155	7,9
<i>Boosting</i> , pérdida absoluta	0,8673	0,8624	0,8673	0,8972	0,6
<i>Boosting</i> optimizado (Optuna)	0,8682	0,8938	0,8682	0,8881	0,0
<i>T₁ anclado</i>					
Predictor constante (referencia)	0,6899	0,6899	0,6880	0,6899	0,3
GLM gaussiano, enlace identidad	0,5829	0,5683	0,5609	0,5842	3,8
<i>Boosting</i> , pérdida cuadrática	0,6003	0,5669	0,5547	0,5974	7,6
<i>Boosting</i> , pérdida absoluta	0,5250	0,5397	0,5266	0,5283	0,0
<i>Boosting</i> optimizado (Optuna)	0,5316	0,5232	0,5367	0,5307	1,6
<i>T₂ incremento</i>					
Predictor constante (referencia)	0,3938	0,3938	0,3936	0,3938	0,1
GLM gaussiano, enlace identidad	0,4375	0,4196	0,3969	0,4289	9,3
<i>Boosting</i> , pérdida cuadrática	0,4435	0,4006	0,3831	0,4383	13,6
<i>Boosting</i> , pérdida absoluta	0,3684	0,3679	0,3729	0,3676	0,2
<i>Boosting</i> optimizado (Optuna)	0,3513	0,3552	0,3654	0,3516	0,0

La lectura de la tabla A.1 confirma la predicción teórica formulada en la sección 3.2.7. Para el objetivo T_2 , el *boosting* con pérdida cuadrática mejora un 13,6 por ciento al eliminar atípicos y el modelo lineal generalizado un 9,3 por ciento, mientras que el *boosting* con pérdida absoluta mejora únicamente un 0,2 por ciento. Para el objetivo T_1 anclado las cifras son, respectivamente, 7,6, 3,8 y 0,0 por ciento.

Puesto que el modelo finalmente seleccionado emplea pérdida absoluta, el tratamiento de atípicos resulta redundante, y por ello los resultados del capítulo 4 se presentan sin aplicar ninguno.

Apéndice B

Optimización de hiperparámetros

La tabla B.1 recoge el espacio de búsqueda explorado por el estimador de Parzen (3.29) y la configuración ganadora para cada uno de los tres objetivos. La búsqueda emplea treinta ensayos. La validación interna es estrictamente temporal: los parámetros se estiman con la temporada 2023 y las configuraciones se comparan según el error obtenido sobre la temporada 2024, de modo que la de 2025 no interviene en ningún momento en la elección, conforme al esquema de tres muestras descrito en la sección 3.2.8. Fijada la configuración ganadora, el modelo se reajusta sobre las dos temporadas de ajuste completas antes de evaluarlo sobre 2025.

CUADRO B.1: Espacio de búsqueda explorado por el estimador de Parzen y configuración ganadora para cada objetivo. Treinta ensayos, con ajuste sobre 2023 y validación sobre 2024, de modo que 2025 no interviene en la elección.

Hiperparámetro	Espacio explorado	T_1 dir.	T_1 anc.	T_2
Número de árboles M	entero en [300, 1500], paso 100	1100	400	600
Tasa de aprendizaje η	real en [0,01, 0,12], escala logarítmica	0,0356	0,0108	0,0102
Número de hojas J	entero en [15, 127], escala logarítmica	23	33	15
Tamaño mínimo de hoja	entero en [5, 80]	21	61	24
Submuestreo de filas	real en [0,6, 1,0]	0,6026	0,6019	0,7638
Submuestreo de columnas	real en [0,5, 1,0]	0,7103	0,5136	0,7804
Penalización λ	real en [0,001, 20], escala logarítmica	4,4807	2,8187	2,5084
MAE de validación	sobre 2024	1,3811	0,6400	0,4185

Las configuraciones ganadoras son informativas por sí mismas. Para los tres objetivos la búsqueda reduce la tasa de aprendizaje respecto del valor de partida, de 0,030 a 0,011 en el objetivo T_1 anclado y a 0,010 en el objetivo T_2 ; para este último reduce además el número de hojas de 31 a 15. Ambos cambios apuntan en la misma dirección, la de un modelo más regularizado, lo que resulta coherente con el diagnóstico de sobreajuste que arrojan los cocientes entre error de prueba y error de ajuste de la tabla I.1.

Apéndice C

Detalle de los valores atípicos

Porcentajes exactos que sustentan la figura 4.2, para las diez variables con mayor proporción de valores marcados como atípicos por las dos reglas de la sección 3.2.7.

CUADRO C.1: Diez variables con mayor proporción de valores atípicos. RIC aplica la regla del rango intercuartílico (3.28) y $3s$ la de tres desviaciones típicas. La discrepancia entre ambas columnas mide asimetría, no error de medida.

Variable	Aus. (%)	RIC (%)	$3s$ (%)	γ_1
s1_compound_ord	0,0	31,7	0,0	-0,01
pressure	10,4	22,4	7,7	-2,32
vueltas_trafico_t	0,0	16,7	3,0	3,64
tren_parada	9,6	11,7	2,7	3,20
hueco_libre_delante	9,6	11,2	7,4	3,16
delta1_t	4,5	10,4	1,9	0,20
pend_t	9,0	10,1	1,6	-4,96
s1_deg	0,0	9,7	2,2	-1,96
s2_tyre_age_start	0,0	9,6	2,4	2,87
T2_incremento	9,9	9,3	2,0	0,47

Apéndice D

Valores exactos de la interpretación

Se recogen aquí los valores numéricos que sustentan las figuras de la sección 4.3.7: la importancia global según los valores de Shapley (3.23) para cada objetivo, y la tabla de coeficientes tipificados del modelo lineal generalizado para el objetivo T_2 .

CUADRO D.1: Doce variables de mayor importancia según los valores de Shapley para el objetivo T_1 anclado, medida como $\frac{1}{n} \sum_i |\varphi_j(x_i)|$ sobre la muestra de prueba. La magnitud está en segundos.

Variable	Descripción	Importancia (s)
s2_start_lap	Vuelta de carrera en que arranca el stint 2	0,2069
lap_number	Vuelta de carrera (aproxima el combustible)	0,0724
s1_pace	Mediana de vueltas limpias del primer stint (s)	0,0646
s1_len	Longitud del primer stint (vueltas)	0,0557
fp_laps	Vueltas completadas en libres	0,0461
s1_last3	Media de las tres últimas del primer stint (s)	0,0424
fp_pace	Mediana del tiempo por vuelta en libres (s)	0,0417
qual_best	Mejor tiempo en clasificación (s)	0,0397
s1_pace_rel	Ritmo del stint 1 menos el mediano de la carrera (s)	0,0369
air_temperature	Temperatura del aire (°C)	0,0318
s1_traffic	Fracción de vueltas del stint 1 con coche a < 1,5 s	0,0311
s1_gap_leader_end	Diferencia con el líder al parar (s)	0,0277

CUADRO D.2: Doce variables de mayor importancia según los valores de Shapley para el objetivo T_2 incremento, medida como $\frac{1}{n} \sum_i |\varphi_j(x_i)|$ sobre la muestra de prueba. La magnitud está en segundos.

Variable	Descripción	Importancia (s)
dev_t	Desviación de la vuelta t respecto al ancla (s)	0,1938
delta1_t	Incremento de la última vuelta, $y_t - y_{t-1}$ (s)	0,0836
lap_number	Vuelta de carrera (aproxima el combustible)	0,0427
pend_t	Pendiente local del stint 2 hasta t (s/vuelta)	0,0407
s2_start_lap	Vuelta de carrera en que arranca el stint 2	0,0280
y_t	Tiempo observado en la vuelta t (s)	0,0183
media3_t	Media de las tres últimas vueltas observadas (s)	0,0176
intervalo_t	Intervalo al coche precedente en t (s)	0,0171
s1_tyre_age_end	Edad del neumático al parar (vueltas)	0,0144
s1_pace_rel	Ritmo del stint 1 menos el mediano de la carrera (s)	0,0117
tyre_age	Edad del neumático en la vuelta predicha	0,0108
track_temperature	Temperatura del asfalto (°C)	0,0106

CUADRO D.3: Diez coeficientes de mayor magnitud del GLM gaussiano para el objetivo T_2 incremento. Las variables están tipificadas, de modo que $\hat{\beta}_j$ es el cambio esperado en el objetivo, en segundos, ante un aumento de una desviación típica en el regresor, manteniendo los demás constantes.

Variable	Descripción	$\hat{\beta}_j$	$\hat{e}e$	z
y_t	Tiempo observado en la vuelta t (s)	-4,7892	0,4571	-10,5
media3_t	Media de las tres últimas vueltas observadas (s)	3,9070	0,4025	9,7
s1_last3	Media de las tres últimas del primer stint (s)	1,1073	0,2071	5,3
qual_best	Mejor tiempo en clasificación (s)	-0,2582	0,0915	-2,8
dev_t	Desviación de la vuelta t respecto al ancla (s)	-0,2451	0,0194	-12,6
s2_start_lap	Vuelta de carrera en que arranca el stint 2	-0,2293	0,0204	-11,2
s1_pace_rel	Ritmo del stint 1 menos el mediano de la carrera (s)	-0,1726	0,0160	-10,8
abrasion	Abrasividad del asfalto (escala 1-5)	-0,1625	0,0190	-8,6
s1_compound_ord	Compuesto del primer stint (1 blando-3 duro)	-0,1156	0,0201	-5,8
tyre_stress	Exigencia estructural sobre el neumático (1-5)	0,0985	0,0193	5,1

Apéndice E

Dependencia de los valores de Shapley

Figura que sustenta la discusión de la sección 4.3.7.

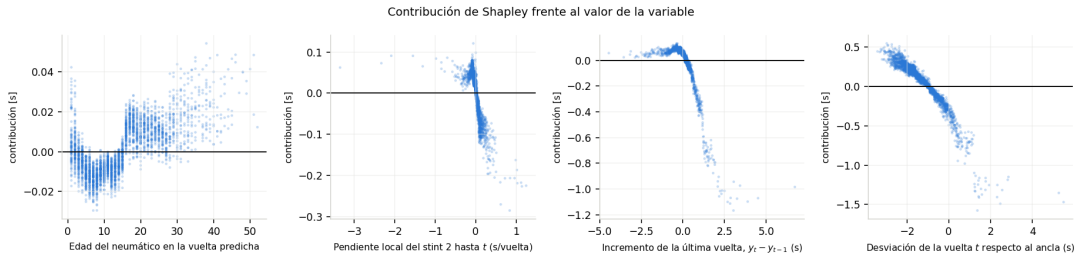


FIGURA E.1: Contribución de Shapley frente al valor de la variable, para las cuatro variables más importantes del objetivo T_2 . La pendiente negativa es la reversión a la media.

Apéndice F

Definición de las variables explicativas

Se recogen las cincuenta y cuatro variables explicativas construidas en la sección 4.3.2, agrupadas por bloque de información, con su definición y el porcentaje de valores ausentes sobre las 17 093 observaciones de la tabla de modelado.

CUADRO F.1: Variables explicativas, agrupadas por bloque de información. La columna *Aus.* recoge el porcentaje de valores ausentes y $|\gamma_1|$ el valor absoluto del coeficiente de asimetría.

Variable	Definición	Aus. (%)	$ \gamma_1 $
<i>Circuito</i>			
abrasion	Abrasividad del asfalto (escala 1–5)	0,0	0,70
tyre_stress	Exigencia estructural sobre el neumático (1–5)	0,0	0,26
overtaking_difficulty	Dificultad de adelantamiento (1–5)	0,0	0,32
aero_load	Carga aerodinámica del trazado (1–5)	0,0	0,23
race_total_laps	Vueltas programadas de la carrera	0,0	0,17
is_sprint	Indicadora de fin de semana al esprint	0,0	1,02
<i>Clasificación</i>			
qual_best	Mejor tiempo en clasificación (s)	0,5	0,57
qual_gap_pole	Diferencia con la pole (s)	0,5	6,80
qual_rank	Posición de salida en parrilla	0,5	0,09
delta_pos_stint1	Posiciones ganadas en el primer stint	0,5	0,09
<i>Entrenamientos libres</i>			
fp_pace	Mediana del tiempo por vuelta en libres (s)	4,7	1,21
fp_deg	Pendiente tiempo–edad en libres (s/vuelta)	6,7	5,53
fp_laps	Vueltas completadas en libres	4,7	0,43
<i>Primer stint</i>			
s1_pace	Mediana de vueltas limpias del primer stint (s)	0,0	0,20
s1_last3	Media de las tres últimas del primer stint (s)	0,0	0,22
s1_deg	Pendiente tiempo–edad en el primer stint (s/vuelta)	0,0	1,96
s1_len	Longitud del primer stint (vueltas)	0,0	0,78
s1_compound_ord	Compuesto del primer stint (1 blando–3 duro)	0,0	0,01

Variable	Definición	Aus. (%)	$ \gamma_1 $
s1_traffic	Fracción de vueltas del stint 1 con coche a $< 1,5$ s	0,0	0,35
s1_traffic_2s	Fracción de vueltas del stint 1 con coche a < 2 s	0,0	0,70
s1_gap_leader_end	Diferencia con el líder al parar (s)	9,6	0,88
s1_pos_end	Posición en carrera al parar	0,0	0,31
s1_pace_rel	Ritmo del stint 1 menos el mediano de la carrera (s)	0,0	0,13
s1_tyre_age_end	Edad del neumático al parar (vueltas)	0,0	0,78
s1_interval_ahead_end	Intervalo al coche precedente al parar (s)	9,8	4,46
s1_tren_medio	Coches encadenados por delante, media del stint 1	7,9	1,44
<i>Plan del segundo stint</i>			
s2_compound_ord	Compuesto del segundo stint (1 blando–3 duro)	0,0	1,24
s2_cnumber	Compuesto absoluto de la gama (C1–C6)	0,0	0,15
s2_tyre_age_start	Edad del juego al montarlo (vueltas)	0,0	2,87
s2_start_lap	Vuelta de carrera en que arranca el stint 2	0,0	0,82
compound_change	Salto de compuesto relativo entre stints	0,0	1,07
cnumber_change	Salto de compuesto absoluto entre stints	0,0	1,07
<i>Pelotón al parar</i>			
densidad_2s	Coches a menos de 2 s al parar	9,6	1,01
densidad_5s	Coches a menos de 5 s al parar	9,6	0,86
coches_delante_10s	Coches por delante en 10 s al parar	9,6	1,09
coches_ventana_reincorporacion	Coches en la ventana de reincorporación	9,6	0,97
hueco_libre_delante	Hueco libre por delante al parar (s)	9,6	3,16
pos_estimada_tras_parar	Posición estimada tras la parada	9,6	0,33
tren_parada	Coches encadenados por delante al parar	9,6	3,20
<i>Meteorología</i>			
air_temperature	Temperatura del aire (°C)	10,4	0,34
track_temperature	Temperatura del asfalto (°C)	10,4	0,27
humidity	Humedad relativa (%)	10,4	0,13
wind_speed	Velocidad del viento (m/s)	10,4	0,72
pressure	Presión atmosférica (hPa)	10,4	2,32
<i>Posición en el stint</i>			
h	Vuelta del segundo stint (horizonte)	0,0	0,83
tyre_age	Edad del neumático en la vuelta predicha	0,0	0,80
lap_number	Vuelta de carrera (aproxima el combustible)	0,0	0,28
<i>Estado observado hasta t (solo $T2$)</i>			
y_t	Tiempo observado en la vuelta t (s)	0,0	0,23
dev_t	Desviación de la vuelta t respecto al ancla (s)	0,0	0,75
media3_t	Media de las tres últimas vueltas observadas (s)	0,0	0,22

Variable	Definición	Aus. (%)	$ \gamma_1 $
<code>pend_t</code>	Pendiente local del stint 2 hasta t (s/vuelta)	9,0	4,96
<code>delta1_t</code>	Incremento de la última vuelta, $y_t - y_{t-1}$ (s)	4,5	0,20
<code>trafico_medio_t</code>	Fracción de vueltas en tráfico hasta t	0,0	0,86
<code>intervalo_t</code>	Intervalo al coche precedente en t (s)	9,1	3,64
<code>vueltas_trafico_t</code>	Vueltas consecutivas en tráfico hasta t	0,0	3,64

Apéndice G

Descriptivo univariante completo

Se recoge el descriptivo univariante de las tres definiciones de la variable de respuesta y de las cincuenta y cuatro variables explicativas construidas en la sección 4.3.2, agrupadas por bloque de información. Los coeficientes γ_1 y γ_2 son los definidos en (3.26).

CUADRO G.1: Descriptivo univariante. $\bar{x}$ media, s desviación típica, γ_1 asimetría y γ_2 exceso de curtosis.

Variable	$\bar{x}$	s	mín.	mediana	máx.	γ_1	γ_2
T1_directo	88,2	9,9	69,2	87,3	117,5	0,23	-0,69
T1_anclado	-1,142	0,884	-4,913	-1,159	6,753	0,75	4,00
T2_incremento	0,020	0,672	-7,221	-0,001	7,287	0,47	13,05
abrasion	2,628	0,841	1,000	3,000	5,000	0,70	0,64
tyre_stress	3,145	1,228	1,000	3,000	5,000	0,26	-1,26
overtaking_difficulty	2,974	1,116	1,000	3,000	5,000	0,32	-0,55
aero_load	3,160	1,215	1,000	3,000	5,000	-0,23	-0,93
race_total_laps	60,9	7,9	44,0	58,0	79,0	0,17	-1,25
is_sprint	0,273	0,445	0,000	0,000	1,000	1,02	-0,96
qual_best	83,1	10,4	64,0	82,4	123,2	0,57	0,49
qual_gap_pole	1,184	1,838	0,000	0,870	23,384	6,80	60,60
qual_rank	9,990	5,711	1,000	10,000	20,000	0,09	-1,19
delta_pos_stint1	-1,935	3,538	-16,000	-1,000	15,000	0,09	2,28
fp_pace	92,1	13,9	69,6	90,0	168,0	1,21	2,69
fp_deg	-0,164	2,176	-10,567	-0,352	35,584	5,53	73,73
fp_laps	27,6	11,5	1,0	25,0	58,0	0,43	-0,40
s1_pace	89,3	10,1	70,2	87,8	116,2	0,20	-0,70
s1_last3	89,5	10,1	70,8	88,1	116,0	0,22	-0,71
s1_deg	-0,050	0,147	-0,998	-0,029	0,446	-1,96	8,32
s1_len	17,0	8,1	5,0	16,0	55,0	0,78	0,61
s1_compound_ord	1,972	0,562	1,000	2,000	3,000	-0,01	0,15
s1_traffic	0,445	0,320	0,000	0,389	1,000	0,35	-1,08
s1_traffic_2s	0,683	0,312	0,000	0,750	1,000	-0,70	-0,66
s1_gap_leader_end	22,5	16,5	0,0	20,1	87,1	0,88	0,62
s1_pos_end	8,080	4,909	1,000	7,000	20,000	0,31	-0,94
s1_pace_rel	-0,065	0,633	-2,427	-0,003	2,797	-0,13	0,80
s1_tyre_age_end	18,3	8,6	4,0	17,0	56,0	0,78	0,69
s1_interval_ahead_end	3,208	4,111	0,000	1,857	56,752	4,46	35,11
s1_tren_medio	2,851	2,779	0,000	2,000	15,000	1,44	1,99

Variable	$\bar{x}$	s	mín.	mediana	máx.	γ_1	γ_2
s2_compound_ord	2,576	0,641	1,000	3,000	3,000	-1,24	0,34
s2_cnumber	2,604	0,934	1,000	3,000	5,000	0,15	-0,35
s2_tyre_age_start	1,601	1,292	1,000	1,000	14,000	2,87	10,71
s2_start_lap	23,4	9,2	9,0	22,0	59,0	0,82	0,58
compound_change	0,605	0,833	-2,000	1,000	2,000	-1,07	0,27
cnumber_change	-0,646	0,960	-3,000	-1,000	3,000	1,07	0,89
densidad_2s	1,398	1,409	0,000	1,000	8,000	1,01	0,61
densidad_5s	3,707	2,733	0,000	3,000	13,000	0,86	0,12
coches_delante_10s	3,197	2,623	0,000	3,000	13,000	1,09	1,00
coches_ventana_	1,734	1,517	0,000	1,000	9,000	0,97	1,21
reincorporacion							
hueco_libre_delante	10,4	25,3	0,1	2,4	99,0	3,16	8,19
pos_estimada_tras_parar	13,0	5,5	2,0	14,0	21,0	-0,33	-1,17
tren_parada	0,957	2,067	0,000	0,000	13,000	3,20	11,30
air_temperature	24,7	4,9	14,3	26,5	33,4	-0,34	-0,95
track_temperature	36,9	9,0	17,3	36,4	51,5	-0,27	-0,70
humidity	49,6	16,8	17,0	50,0	78,0	-0,13	-1,07
wind_speed	1,741	0,783	0,300	1,700	5,000	0,72	0,54
pressure	978,5	62,7	779,2	1.006,3	1.020,8	-2,32	4,45
h	13,7	9,4	1,0	12,0	54,0	0,83	0,34
tyre_age	14,3	9,4	1,0	13,0	58,0	0,80	0,33
lap_number	36,1	12,8	9,0	35,0	72,0	0,28	-0,54
y_t	88,2	9,9	69,2	87,3	117,5	0,23	-0,69
dev_t	-1,142	0,884	-4,913	-1,159	6,753	0,75	4,00
media3_t	88,2	9,9	69,4	87,2	116,6	0,22	-0,70
pend_t	-0,004	0,171	-4,940	-0,005	2,442	-4,96	128,21
delta1_t	0,017	0,642	-9,674	-0,002	6,861	0,20	23,37
trafico_medio_t	0,314	0,324	0,000	0,210	1,000	0,86	-0,47
intervalo_t	4,438	5,699	0,000	2,613	87,111	3,64	22,27
vueltas_trafico_t	1,592	3,990	0,000	0,000	39,000	3,64	15,48

Apéndice H

Descriptivo bivalente completo

Medias por tramo (3.27) de cada variable explicativa, con su intervalo al noventa y cinco por ciento, para los dos objetivos analizados en la sección 4.3.3.

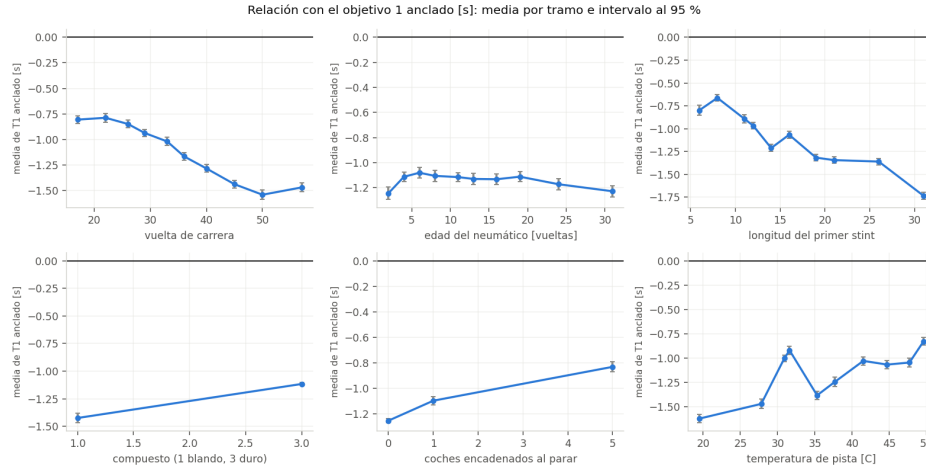


FIGURA H.1: Media del objetivo T_1 anclado por tramo de cada variable explicativa, con intervalo al noventa y cinco por ciento. Tramos definidos por deciles.

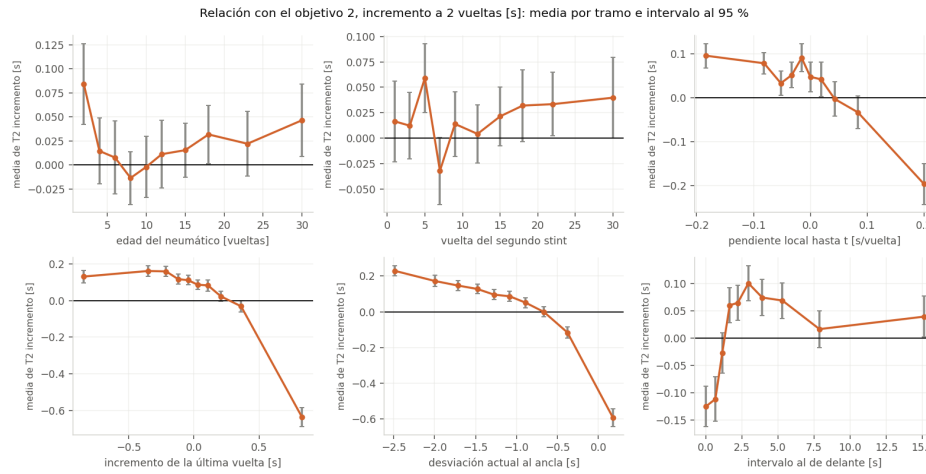


FIGURA H.2: Media del objetivo T_2 por tramo de cada variable explicativa, con intervalo al noventa y cinco por ciento.

Apéndice I

Capacidad predictiva sobre los tres objetivos

Valores exactos que sustentan la discusión de la sección 4.3.6.

CUADRO I.1: Capacidad predictiva sobre los tres objetivos. Ajuste con las temporadas 2023 y 2024, prueba sobre 2025. Las columnas de MAE y RMSE están en segundos; la última es el cociente entre el MAE de prueba y el de ajuste, que resume el sobreajuste.

Modelo	MAE ajuste	MAE prueba	RMSE prueba	R^2 prueba	Cociente
<i>T₁ directo (tiempo por vuelta, en segundos)</i>					
Predictores de referencia					
Predictor constante	8,7418	7,9986	9,0857	-0,0352	0,91
Constante por piloto y circuito	0,5836	2,5428	3,2523	0,8674	4,36
Persistencia	1,1784	1,3405	1,5210	0,9710	1,14
Modelos ajustados					
GLM gaussiano, enlace identidad	0,5049	0,5829	0,7743	0,9925	1,15
GLM Gamma, enlace logarítmico	0,5033	0,7444	0,9303	0,9891	1,48
<i>Boosting</i> , pérdida cuadrática	0,2893	0,9146	1,4688	0,9729	3,16
<i>Boosting</i> , pérdida absoluta	0,2892	0,8673	1,3532	0,9770	3,00
<i>Boosting</i> optimizado	0,2778	0,8682	1,3628	0,9767	3,13
<i>T₁ anclado (desviación respecto al primer stint)</i>					
Predictores de referencia					
Predictor constante	0,6597	0,6899	0,8768	-0,0508	1,05
Constante por piloto y circuito	0,4405	0,7448	0,9443	-0,2189	1,69
Persistencia	1,1784	1,3405	1,5210	-2,1622	1,14
Modelos ajustados					
GLM gaussiano, enlace identidad	0,5049	0,5829	0,7743	0,1806	1,15
<i>Boosting</i> , pérdida cuadrática	0,2851	0,6003	0,7888	0,1495	2,11
<i>Boosting</i> , pérdida absoluta	0,2870	0,5250	0,7104	0,3102	1,83
<i>Boosting</i> optimizado	0,3688	0,5316	0,7179	0,2956	1,44
<i>T₂ (incremento a dos vueltas)</i>					
Predictores de referencia					
Predictor constante	0,4390	0,3938	0,6308	-0,0004	0,90
Constante por piloto y circuito	0,4251	0,4064	0,6401	-0,0300	0,96
Persistencia	0,4391	0,3934	0,6307	-0,0000	0,90
Último incremento observado	0,6983	0,6232	0,9821	-1,6128	0,89
Modelos ajustados					
GLM gaussiano, enlace identidad	0,3971	0,4375	0,6027	0,0868	1,10
<i>Boosting</i> , pérdida cuadrática	0,2539	0,4435	0,6061	0,0766	1,75
<i>Boosting</i> , pérdida absoluta	0,2735	0,3684	0,5526	0,2323	1,35
<i>Boosting</i> optimizado	0,3412	0,3513	0,5459	0,2510	1,03

Bibliografía

- [1] Akiba, T., Sano, S., Yanase, T., Ohta, T. y Koyama, M. (2019). *Optuna: a next-generation hyper-parameter optimization framework*. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2623–2631.
- [2] Anandatama, D. H. y Gamayanto, I. (2025). *Optimizing F1 tyre performance prediction with SVC, XGBoost, and Optuna for Dutch GP 2022*. Journal of Applied Informatics and Computing, 9(6), 3110–3117.
- [3] Azevedo, A. y Santos, M. F. (2008). *KDD, SEMMA and CRISP-DM: a parallel overview*. IADIS European Conference on Data Mining, 182–185.
- [4] Bekker, J. y Lotz, W. (2009). *Planning Formula One race strategies using discrete-event simulation*. Journal of the Operational Research Society, 60(7), 952–961.
- [5] Bergmeir, C. y Benítez, J. M. (2012). *On the use of cross-validation for time series predictor evaluation*. Information Sciences, 191, 192–213.
- [6] Bergstra, J., Bardenet, R., Bengio, Y. y Kégl, B. (2011). *Algorithms for hyper-parameter optimization*. Advances in Neural Information Processing Systems, 24, 2546–2554.
- [7] Bergstra, J. y Bengio, Y. (2012). *Random search for hyper-parameter optimization*. Journal of Machine Learning Research, 13, 281–305.
- [8] Breiman, L., Friedman, J. H., Olshen, R. A. y Stone, C. J. (1984). *Classification and regression trees*. Wadsworth.
- [9] Breiman, L. (1996). *Bagging predictors*. Machine Learning, 24(2), 123–140.
- [10] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. y Wirth, R. (2000). *CRISP-DM 1.0: step-by-step data mining guide*. CRISP-DM Consortium.
- [11] Chen, T. y Guestrin, C. (2016). *XGBoost: a scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
- [12] Choo, C. L. W. (2015). *Real-time decision making in motorsports: analytics for improving professional car race strategy*. Tesis de máster, Massachusetts Institute of Technology.
- [13] Efron, B. y Morris, C. (1975). *Data analysis using Stein’s estimator and its generalizations*. Journal of the American Statistical Association, 70(350), 311–319.
- [14] Schaefer, P. (2024). *FastF1: Python package for accessing and analyzing Formula 1 results, schedules, timing data and telemetry*. <https://docs.fastf1.dev>
- [15] Friedman, J. H. (2001). *Greedy function approximation: a gradient boosting machine*. The Annals of Statistics, 29(5), 1189–1232.
- [16] Frisch, R. y Waugh, F. V. (1933). *Partial time regressions as compared with individual trends*. Econometrica, 1(4), 387–401.
- [17] Gelman, A. y Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

- [18] Heilmeier, A., Graf, M. y Lienkamp, M. (2018). *A race simulation for strategy decisions in circuit motorsports*. 21st International Conference on Intelligent Transportation Systems (ITSC), 2986–2993. IEEE.
- [19] Heilmeier, A., Graf, M., Betz, J. y Lienkamp, M. (2020). *Application of Monte Carlo methods to consider probabilistic effects in a race simulation for circuit motorsport*. Applied Sciences, 10(12), 4229.
- [20] Hoerl, A. E. y Kennard, R. W. (1970). *Ridge regression: biased estimation for nonorthogonal problems*. Technometrics, 12(1), 55–67.
- [21] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. y Liu, T.-Y. (2017). *LightGBM: a highly efficient gradient boosting decision tree*. Advances in Neural Information Processing Systems, 30, 3146–3154.
- [22] Koenker, R. y Bassett, G. (1978). *Regression quantiles*. Econometrica, 46(1), 33–50.
- [23] Liu, X. y Fotouhi, A. (2020). *Formula-E race strategy development using artificial neural networks and Monte Carlo tree search*. Neural Computing and Applications, 32(18), 15191–15207.
- [24] Lovell, M. C. (1963). *Seasonal adjustment of economic time series and multiple regression analysis*. Journal of the American Statistical Association, 58(304), 993–1010.
- [25] Lundberg, S. M. y Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. Advances in Neural Information Processing Systems, 30, 4765–4774.
- [26] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. y Lee, S.-I. (2020). *From local explanations to global understanding with explainable AI for trees*. Nature Machine Intelligence, 2(1), 56–67.
- [27] McCullagh, P. y Nelder, J. A. (1989). *Generalized linear models*, 2.^a edición. Chapman and Hall.
- [28] Nelder, J. A. y Wedderburn, R. W. M. (1972). *Generalized linear models*. Journal of the Royal Statistical Society A, 135(3), 370–384.
- [29] Godefroy, B. (2024). *OpenF1: free and open-source API providing real-time and historical Formula 1 data*. <https://openf1.org>
- [30] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. y Gulin, A. (2018). *CatBoost: unbiased boosting with categorical features*. Advances in Neural Information Processing Systems, 31, 6639–6649.
- [31] Sasikumar, A., Leema, A. A. y Balakrishnan, P. (2025). *Data-driven pit stop decision support for Formula 1 using deep learning models*. Frontiers in Artificial Intelligence, 8, 1673148.
- [32] SAS Institute Inc. (2017). *SAS Enterprise Miner 14.3: reference help*. SAS Institute Inc., Cary, NC.
- [33] Seber, G. A. F. y Lee, A. J. (2003). *Linear regression analysis*, 2.^a edición. Wiley.
- [34] Shapley, L. S. (1953). *A value for n-person games*. En *Contributions to the Theory of Games*, vol. II, 307–317. Princeton University Press.
- [35] Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society B, 58(1), 267–288.
- [36] Todd, J., Jiang, J., Russo, A., Winkler, S., Sale, S., McMillan, J. y Rago, A. (2025). *Explainable time series prediction of tyre energy in Formula One race strategy*. Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing, 1082–1089.
- [37] Tulabandhula, T. y Rudin, C. (2014). *Tire changes, fresh air, and yellow flags: challenges in predictive analytics for professional racing*. Big Data, 2(2), 97–112.