quantil

BOLETÍN DE MATEMÁTICAS APLICADAS A LA INDUSTRIA

Predicción de precios del parque automotor en Colombia

Santiago Pulgarín*, Andres Galeano†,Juan Pablo Salas‡, Juan Esteban Segura§

No. 6
27 de junio de 2024

Resumen

El parque automotor mundial entre 2020 y 2022 sufrió una pérdida en la capacidad productiva debido a la escasez de microchips, generando un impacto significativo sobre los precios de vehículos nuevos y usados. Poco a poco, esta escasez se ha superado y hemos retornado a la normalidad. El impacto sobre los precios resulta difícil de rastrear sin técnicas de analítica apropiadas, que permitan mapear los precios del mercado actuales utilizando una submuestra del universo.

Con este objetivo en mente, Quantil desarrolló una herramienta de proyección de los precios actuales (*"nowcasting"*), el cual permite poblar la Guía de Valores, referencia en el país para los precios del parque automotor, incorporando información pública de los portales de venta de vehículos usados.

Los resultados obtenidos permiten poblar el grueso del parque automotor con precios dinámicos que responden a la realidad económica y de mercado del país. Asimismo, en pruebas posteriores se observó una alta cercanía con el precio efectivo de venta de los concesionarios de autos usados.

La Guía de Valores ha sido históricamente un referente nacional tanto para el sector asegurador como para el público general, partiendo por el aseguramiento de vehículos y pasando por su uso para compraventa de usados. Sin embargo, 2021 y 2022 presentaron retos muy relevantes en el comportamiento del mercado, por diversas razones que pueden resumirse en (i) incremento de la demanda post-covid, (ii) retos logísticos debido a fallas en varias etapas de la cadena de suministros, particularmente por falta de microchips y, como consecuencias de estos dos, (iii) listas de espera de hasta dos años y (iv) crecimiento acelerado de los precios de vehículos nuevos.

Acompañado de esta situación, la Guía de Valores entró en 2022 en un proceso de transición, en el cual se da un cambio de proveedor y metodología, que busca implementar elementos de analítica de datos que permitan mantener actualizados los precios de manera más fiel al mercado, incorporando la información observada para crear precios representativos de estos.

Como consecuencia de este proceso, en agosto del 2022 se genera la primera actualización masiva de los precios de vehículos del mercado automotor colombiano en la Guía de Valores, observando crecimientos importantes en la mayoría de segmentos - del orden del 25 % -, siendo los vehículos más antiguos los que presentan los mayores cambios, con cambios porcentuales del orden de 50 %, y para vehículos del 2000-2005 hasta del 100 %.

Los movimientos observados están alineados con los movimientos del mercado a nivel mundial, los cuales habían exhibido crecimientos promedios de 50 % desde el punto más bajo (junio 2020) hasta el más alto (enero 2022). En Colombia, sin embargo, este movimiento no se había visto reflejado en la Guía hasta la inclusión de la nueva metodología. Asimismo, los precios de vehículos nuevos desde agosto 2022 a junio 2023 continuaron incrementando cerca de 8 %, lo que contribuyó a futuros incrementos en los precios de usados. Asimismo, se presentó una situación atípica en el mercado, dada la longitud de las listas de espera: los precios de vehículos usados último modelo de bajo kilometraje podían superar los precios de nuevos, dado que los primeros tienen disponibilidad inmediata, frente a listas de más de un año en los segundos.

*BA en Ingeniería Biomédica e Ingeniería Industrial, Universidad de los Andes
†BA en Economía y Administración de Empresas, MSc. en Economía, Universidad de los Andes
‡BA en Física, Universidad de los Andes
§BA en Economía, Universidad de los Andes

Boletín de Matemáticas Aplicadas a la Industria es una publicación de Quantil S.A.S.

A diciembre 2023, en cambio, ha culminado la estabilización y se cumplen ocho meses de caídas sostenidas, con un acumulado del orden del 10 % en el precio de vehículos usados. Asimismo, se presentan caídas importantes en las ventas de nuevos y usados, tasas de interés altas y altos niveles de inventarios, para el grueso de marcas.

El presente boletín se presenta a modo informativo para la industria, a fin de exponer la metodología de estimación de precios y las posibles aplicaciones extendidas de dicha metodología.

1. Marco conceptual

La actualización periódica de precios en la Guía de Valores requiere de un esquema de automatización de alto nivel, pero requiere, por su exposición en el mercado, de un alto grado de supervisión.

En su mayoría, los datos son obtenidos de la web (portales de ventas de usados) y de fuentes aliadas, como marcas y concesionarios de autos. Estos datos contienen información bruta sin el grado de pre-procesamiento requerido para poder construir un modelo de analítica.

1.1. Pre-procesamiento

El pre-procesamiento de la información inicia con una limpieza de los datos, buscando unificar los formatos provenientes de fuentes diferentes. Así pues, el primer paso de esta limpieza consiste en seleccionar las columnas relevantes de la información extraída, además de añadir algunos campos adicionales para que se pueda contar con la misma información que en versiones anteriores. En este sentido, el conjunto de datos mínimo obtenido de las fuentes es:

- Precio.
- Texto del vehículo (Mazda 2 Touring, por ejemplo).
- Año modelo del vehículo.
- Kilometraje.
- Transmisión (automática, mecánica u otra).
- Combustible (Gasolina, Diésel, Eléctrico, Híbrido, etc.).
- Cilindraje.

Asimismo, apoyándose en la información contenida en la Guía de Valores, una vez se asocian los vehículos a los vehículos registrados en la Guía, se obtiene información adicional, como potencial número de airbags, presencia de elementos de lujo, como asientos de cuero, sunroof, entre otros.

1.2. Detección de outliers

Después de la limpieza de los diferentes campos, se emplea un Isolation Forest, un modelo basado en árboles que se encarga de detectar anomalías o outliers, para llegar a una muestra de datos limpia (8)(10)(6). Así, se filtrarán aquellos datos que el modelo pueda considerar atípicos, reduciendo el riesgo de que estos vehículos afecten el precio final de los carros tipo.

Posteriormente, se emplea otro método de eliminación de outliers, partiendo de los residuales de una regresión lineal, tomando como variable de respuesta el logaritmo natural del precio de los vehículos. De esta manera, al observar la distancia que hay entre los precios y la predicción, se pueden identificar datos atípicos con valores muy alejados de la línea media. Esto permite, en primer lugar, eliminar datos que se alejen mucho del promedio o que se expliquen por factores no observados. Como el objetivo es determinar el precio de los vehículos tipo, esta eliminación de outliers favorece la estabilidad, reduce la varianza y permite una mejor predicción del valor promedio.

1.3. Modelos de analítica

Para la construcción de modelo de analítica se parte de un conjunto de metodologías candidatas, un set de hiperparámetros a revisar y algunos criterios de selección de variables y modelos. En general, se consideran los modelos LGB (light gradient boosting model), regresión lineal y random forest, así como un ensamble de todos los modelos.

1.3.1. Partición en sub-modelos por edad y categoría

Durante la construcción de modelos de analítica se evidencian dificultades cuando se trabaja con un único modelo para todos los años modelo, particularmente por la presencia de vehículos "clásicos" en los años antiguos y la dificultad para diferenciarlos de los no-clásicos. Esto genera un comportamiento forma de "U" del valor de los vehículos, lo que no representa el objetivo de la Guía, ya que los autos con placas de "antiguo" o clásicos no son objeto de la misma cobertura en las compañías aseguradoras. Asimismo, los vehículos de lujo tienen una mayor sensibilidad a ciertas características como potencia y elementos de *comfort* y lujo, por lo que una división de los datos es una alternativa razonable para enfrentar este problema.

Por lo anterior, en primer lugar, se dividen entre (i) motos y (ii) automóviles y camionetas (incluidas pickups). Por su parte, las motos se dividen en categorías de (i) lujo y (ii) no lujo, sin división de año modelo, teniendo en cuenta el bajo volumen de datos para motos antiguas. Por otra parte, los automóviles, camionetas y pickups se dividieron en 6 categorías, generadas por dos divisiones; la primera de ellas separa los vehículos por edad y combustible, dividiéndolos en tres grupos: (i) antiguos sin eléctricos (anteriores al 2000), (ii) recientes sin eléctricos (del 2000 en adelante) y (iii) eléctricos (todos los años); la segunda partición subdivide entre vehículos de (i) lujo y (ii) no lujo.

La división anterior responde a que algunos vehículos antiguos ganan valor en el tiempo en comparación a los vehículos recientes, particularmente para coleccionistas de carros. Asimismo, la regresión lineal no logra ajustarse adecuadamente a

la depreciación de los datos, particularmente para los vehículos de mayor edad. Igualmente, es importante notar que los vehículos antiguos tienen una cantidad significativamente menor de datos, haciéndolos más susceptibles a outliers y aumentando el riesgo de sobre-ajuste.

1.3.2. Selección de hiper-parámetros en los modelos (OPTUNA)

Luego de seleccionar los modelos candidatos, es importante llevar a cabo una calibración los parámetros que emplea cada modelo. Para esto, se hace uso de la metodología OPTUNA, una optimización de hiper-parámetros basada en técnicas de optimización Bayesiana, diseñada para automatizar y robustecer el proceso de predicción de modelos de aprendizaje de máquinas. Utilizando el muestreo *Tree-structured Parzen Estimator* (TPE), una variante de la optimización Bayesiana, OPTUNA crea un modelo probabilístico del rendimiento del objetivo y sugiere nuevos hiper-parámetros basándose en este modelo. Una de las fortalezas clave de este enfoque es ponderar la exploración y explotación: mientras que la exploración busca hiper-parámetros en áreas no evaluadas previamente para descubrir nuevas soluciones potenciales, la explotación se concentra en las regiones de alto desempeño pasado. En particular, este modelo se utiliza para los modelos de boosting implementados, evaluando la tasa de aprendizaje, el número de árboles y la profundidad máxima del árbol. Además de la optimización Bayesiana, OPTUNA utiliza técnicas de "pruning" para explorar eficientemente el espacio de hiper-parámetros (1).

1.3.3. Selección de variables o atributos

En términos generales, la selección de atributos se realiza a través de las importancias de bosques aleatorios iniciales, al tiempo que se evalúa la intuición de negocio y significancia de los coeficientes del modelo de regresión lineal.

1.3.4. Métricas de desempeño de los modelos

Para medir el desempeño de los modelos, se obtienen los errores entre el valor real de la variable de interés con la predicción realizada por los modelos. Se utilizan dos medidas de error diferentes: error absoluto medio (MAE) y error porcentual medio (MAPE).

1.3.5. Modelos evaluados

En primer lugar, en el caso de motos se consideraron tres tipos de modelo: regresión lineal (3)(7), random forest (2) y light GBM(9)(5)(4), siendo este último basado en árboles de decisión. Por otra parte, para los demás vehículos se emplearon dos modelos: (i) un modelo de regresión y (ii) un modelo light GBM.

1.3.6. Selección del modelo ganador

Para seleccionar el modelo ganador, se realiza una comparación de los valores predichos de cada modelo con los precios de oferta observados, obtienendo MAPE (mean average percentage error) en el rango de 4 % a 10 % para todos los modelos y agrupaciones de datos. Es importante recalcar que esta diferencia es sobre ofertas de vehículos y no sobre transacciones reales. Lo anterior difiere de la variable objetivo de la Guía de Valores (el valor promedio del vehículo representativo), por lo que el valor promedio puede encontrarse a una distancia menor que la distancia de cada carro individual. Es importante mencionar que, al final, cada referencia o línea de vehículo puede tener su propio "mejor" modelo. Lo anterior no solo considera elementos de precisión sino también elementos de estabilidad en el tiempo. Este esquema se abstrae un poco de la idoneidad matemática del modelo, pero se enfoca en la aplicabilidad e interpretabilidad de este, integrando elementos como la cantidad de datos observados en un vehículo al momento de definir su idoneidad metodológica (vehículos con uno o unos pocos datos tenderán a ser inestables y/o tener riesgo mayor de sobre-ajuste, en comparación con vehículos con cientos o miles de datos). Adicionalmente, se puede presentar el caso en el que ningún precio se aproxime al precio del mercado del vehículo (vehículos con pocas o nulas observaciones, por ejemplo), por lo que optará por mantener el precio estable hasta conseguir una cantidad de información mayor.

1.3.7. Resultados comparativos

Como primer elemento de comparación entre modelos, se considera la importancia de las variables disponibles, para establecer las de mayor relevancia al momento de predecir los precios de los vehículos. Por ejemplo, en la figura 1 se observa que para los vehículos no-lujo recientes las variables con una mayor importancia son kilometraje, peso, potencia y cilindraje siendo esta superior a un 10 % del modelo completo. Por otra parte, en la figura 2 que presenta la importancia de las variables en el modelo de motos se puede observar como la importancia cambia en comparación al modelo de la figura 1, pues en este caso las variables con mayor importancia son la marca, el cilindraje, el peso y la potencia lo cual involucra algunas de las mismas variables con mayor importancia, pero su nivel de importancia varía.

Por otra parte, en la tabla 1 se puede ver una comparación de los modelos en términos de métricas como lo son el MAE y el MAPE, las cuales resultan tomar valores más bajos para el método de Light GBM dando indicios que este método resulta ser el mejor en una mayor cantidad de casos que los demás métodos. Sin embargo, vale la pena mencionar que en cuanto a la predicción del precio de motos y derivados de ella el método que resulta tener mejores resultados corresponde a la regresión lineal.

Adicionalmente, en la tabla 2 se muestra una comparación de precios entre la guía 321, la cual corresponde a la última

guía en la que se empleo un método de estimación diferente a los métodos propuestos por Quantil S.A.S vigentes desde la guía 322. Particularmente, se destaca que al visualizar los precios del mercado para los vehículos presentados se aproximan mejor a los precios de la guía 322, argumentando a favor del uso del nuevo modelo.

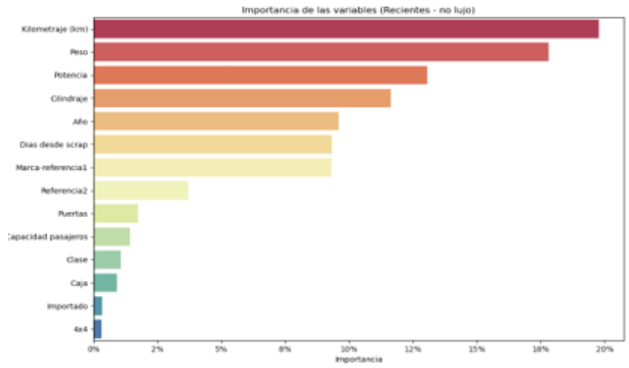

Figura 1. Importancia de variables en el modelo de no lujo - recientes para la estimación de precios de los vehículos.

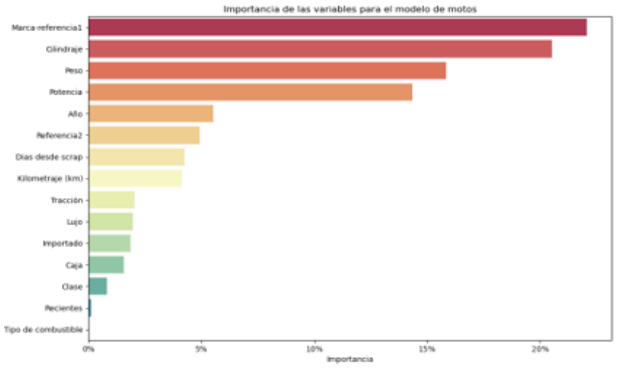

Figura 2. Importancia de variables en el modelo de motos para la estimación de precios.

Modelos	RegLin	RF	Stacking	LightGBM
MAE Train	4,910	4,653	4,234	2,864
MAE Test	4,998	5,290	4,442	3,704
MAPE Train	9.22 %	9.10 %	8.28 %	5.57 %
MAPE Test	9.55 %	10.48 %	8.86 %	7.12 %
Desv. Train	7.52 %	8.79 %	6.97 %	4.61 %
Desv. Test	7.79 %	9.81 %	7.43 %	6.20 %

Cuadro 1. Tabla de comparación de modelos. Cálculos propios. RF: Random Forest; RegLin: Regresion Lineal

Vehículo	Año	Guía 321	Mercado	Guía 322
Ssangyong Actyon	2011	26,200	34,750	32,300
Ford Ranger	2010	37,200	56,420	58,300
Land Rover	2016	247,700	324,000	380,680
Lexus GX	2020	270,800	485,200	449,994

Cuadro 2. Comparación de precios bajo la nueva y antigua metodología y el mercado. Mercado libre y Fasecolda. Fuentes: TuCarro; Fasecolda.

2. Resultados agregados

En las figuras 3 y 4, se puede observar los resultados de la diferencia y/o distancia al comparar los métodos anteriores para la estimación de precios con los precios reales de mercado proporcionados por concesionarios, además de la distancia al comparar los métodos nuevos propuestos por Quantil S.A.S. Particularmente, se puede evidenciar como los métodos anteriores tienen una distribución que tiene una media centrada alrededor de los -7.5M de COP, junto con un 20 % MAPE, mientras que la media de la distribución de distancia de Quantil S.A.S se encuentra más centrada aproximadamente en -1.1M, con un 6 % de MAPE. Además, en la tabla 3 se puede observar la desviación estándar de los datos con los cuales se construyeron los histogramas siendo considerablemente mayor la de los métodos anteriores, pues toma un valor de desviación de 35.6 % comparado con el 4.6 % de desviación para los resultados de Quantil S.A.S.

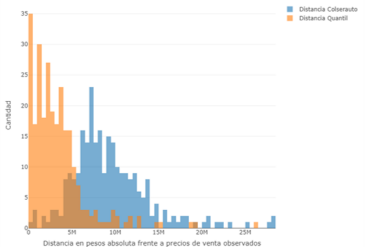

Figura 3. Distancia en pesos absoluta frente a precios observados en el mercado. Fuente: Autonal. Cálculos propios a partir de observaciones de datos de transacción.

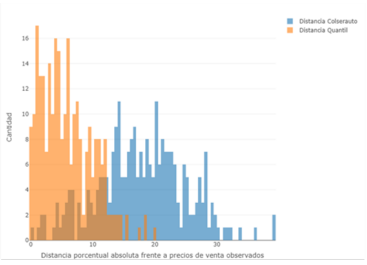

Figura 4. Distancia porcentual absoluta frente a precios observados en el mercado. Fuente: Autonal. Cálculos propios a partir de observaciones de datos de transacción.

Modelo	Anterior	Quantil
Diferencia promedio [COP]	-7,534,546	-1,095,634
Diferencia promedio [%]	20.10 %	6.10 %
Desviación estándar [%]	35.60 %	4.60 %

Cuadro 3. Diferencia de los modelos frente a muestra de precios de transacción. Fuente: Autonal. Cálculos propios a partir de observaciones de datos de transacción.

3. Conclusiones

En términos generales, los modelos construidos para la Guía de Valores, combinando distintas metodologías y etapas de analítica de datos, han permitido a la Guía acercarse considerablemente al mercado, obteniendo una mayor fidelidad del reflejo de la actualidad, al tiempo que le permite actualizar los precios con mayor frecuencia, debido al uso de información disponible en la web. Por otra parte, la incorporación de elementos adicionales de análisis no tradicional, como las revisiones manuales, el monitoreo constante de las predicciones y cambios entre publicaciones de guías y la asignación por línea de vehículos a un modelo específico, permite reducir las debilidades y aprovechar las fortalezas de cada modelo, sin descuidar la cercanía entre los precios del mercado y los precios reflejados en la Guía.

Estos modelos predictivos permiten al cliente estimar información no observada, como los precios de vehículos poco comerciales, al tiempo que permiten estar actualizados en una realidad altamente cambiante, por medio de distintas fuentes de información. Asimismo, a lo largo del estudio se permite inferir las importancias relativas que afectan el valor de los vehículos, lo que permite a las marcas entender mejor las preferencias de los consumidores y la disponibilidad a pagar de sus potenciales clientes.

Referencias

Akiba, T., Sano, S., Yanase, T., Ohta, T., y Koyama, M. (2019). Optuna: A Next-Generation Hyperparameter Optimization Framework. En *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (pp. 2623–2631). New York, NY, USA: Association for Computing Machinery. Descargado de https://doi.org/10.1145/3292500.3330701 doi: 10.1145/3292500.3330701

Čeh, M., Kilibarda, M., Lisec, A., y Bajat, B. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *ISPRS International Journal of Geo-Information*, *7*(5). Descargado de https://www.mdpi.com/2220-9964/7/5/168 doi: 10.3390/ijgi7050168

Draper, N. R., y Smith, H. (1998). *Applied Regression Analysis*. Wiley. Descargado de https://books.google.com.co/books?id=d6NsDwAAQBAJ

Fourkiotis, K. P., y Tsadiras, A. (2023). Comparing Machine Learning Techniques for House Price Prediction. En I. Maglogiannis, L. Iliadis, J. MacIntyre, y M. Dominguez (Eds.), *Artificial intelligence applications and innovations* (pp. 292–303). Cham: Springer Nature Switzerland.

Gao, X., Wang, J., y Yang, L. (2022). An Explainable Machine Learning Framework for Forecasting Crude Oil Price during the COVID-19 Pandemic. *Axioms*, *11*(8). Descargado de https://www.mdpi.com/2075-1680/11/8/374 doi: 10.3390/axioms11080374

Lesouple, J., Baudoin, C., Spigai, M., y Tourneret, J.-Y. (2021). Generalized isolation forest for anomaly detection. *Pattern Recognition Letters*, *149*, 109–119. Descargado de https://www.sciencedirect.com/science/article/pii/S0167865521002063 doi: https://doi.org/10.1016/j.patrec.2021.05.022

Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S., y Boonpou, P. (2018). Prediction of prices for used car by using regression models. En *2018 5th international conference on business and industrial research (icbir)* (pp. 115–119). doi: 10.1109/ICBIR.2018.8391177

Tony Liu, F., Ming Ting, K., y Zhou, Z.-H. (2008). Isolation Forest ICDM08. *Icdm*. Descargado de https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf%0Ahttps://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest

Winky K.O. Ho, B.-S. T., y Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*(1), 48–70. Descargado de https://doi.org/10.1080/09599916.2020.1832558 doi: 10.1080/09599916.2020.1832558

Xu, D., Wang, Y., Meng, Y., y Zhang, Z. (2017). An Improved Data Anomaly Detection Method Based on Isolation Forest. En *2017 10th international symposium on computational intelligence and design (iscid)* (Vol. 2, pp. 287–291). doi: 10.1109/ISCID.2017.202