Análisis y predicción de ventas de motos haciendo uso de la metodología “Customer Value Map” y técnicas de Machine Learning

Proyecto de Grado- Maestría en Ciencia de Datos y Analítica – Universidad EAFIT

Autor: Sandra Marcela Díaz Cordero – smdiazc@eafit.edu.co
Director: Juan David Martínez Vargas - jdmartinev@eafit.edu.co-
Co-Directora: Paola Andrea Vallejo Correa - pvallej3@eafit.edu.co

Mayo 2024

Contenido

1. Resumen

Los problemas de predicción y pronóstico de ventas han sido explorados ampliamente por las compañías dado su poder de generar rentabilidades futuras en la empresa, optimizando los costos al no generar una sobre oferta o escasez del producto.

En particular se han aplicado técnicas estadísticas de series de tiempo clásicas como respuesta al problema planteado, las cuales contemplan las ventas del producto y sus rezagos como las variables explicativas más importantes del modelo de predicción. Sin embargo, este enfoque para resolver los problemas de pronóstico de ventas está obviando la intención de compra del consumidor como variable fundamental para determinar las ventas de una marca específica. De igual manera, para algunos problemas de predicción en series de tiempo, Machine Learning puede ser de gran utilidad, dado que se pueden modelar relaciones no lineales entre las variables y se pueden manejar diferentes variables y características para explicar el fenómeno en la serie de tiempo.

El objetivo de este proyecto es construir un modelo usando técnicas de Machine Learning que permita predecir mensualmente las ventas de motos (vehículos de combustión de 2 ruedas) y participación de mercado de una marca ensambladora del mercado colombiano, incorporando variables de la metodología “Customer Value Map” (teorías del consumidor).

Se espera obtener un modelo que demuestre que los métodos basados en árboles son los más adecuados para predecir las ventas de motos. Adicional, se espera comprobar que la propensión de ganar o perder participación de mercado (unidades vendidas de la marca propia, comparadas con las unidades vendidas de la competencia) es directamente proporcional a la distancia perpendicular de las coordenadas del producto en el “Customer Value Map” a la línea de equilibrio de este mismo mapa. Este resultado, permitirá a la compañía tener pronósticos acertados y conocer el verdadero impacto de la percepción del consumidor en sus ventas.

Palabras claves: Customer Value Map, valor percibido, predicción de ventas, Pronóstico, Machine Learning, Forecasting, Motos

2. Descripción del problema

2.1. Planteamiento del Problema

Las motos se han presentado ante los consumidores como una alternativa de transporte que brinda la comodidad de tener una opción privada y propia de movilizarse, y a la vez es económicamente favorable para el usuario (Sasmita & Darmawan, 2017).

El problema para tratar en el presente proyecto se ubica en la industria de motos nuevas en Colombia; un mercado que mueve alrededor de 700 mil unidades al año[1] (con tendencia creciente año tras año) y en el cual son varias las marcas competidoras que buscan participar activamente en las ventas de la industria (AKT, Auteco, Hero, Honda, Suzuki, UMA, Yamaha).

En una industria en constante crecimiento se han buscado herramientas que permitan comprender el dinamismo del mercado y así definir cuál es la participación de la marca propia en las ventas totales de motos. Así, los esfuerzos de las empresas se han dividido para atacar varios frentes pertinentes: Los equipos de marketing han buscado entender al consumidor y el valor que este le da a su producto (Mohd et al., 2013); los equipos comerciales han tratado de impulsar las ventas aplicando estrategias de precios y monitoreando los movimientos de la competencia; y los equipos de analítica han buscado predecir las ventas en el corto plazo a través de modelaciones econométricas y de técnicas de aprendizaje de máquina. Sin embargo, cada uno de estos han actuado como islas, por lo cual surge el interrogante ¿Se puede predecir las ventas de la compañía haciendo uso de las técnicas que el equipo de marketing y comercial han implementado para entender al consumidor y la competencia?

El problema abordado en este proyecto se encuentra ligado al interrogante anterior. Es decir, se construye un modelo de aprendizaje automático para la predicción de ventas de motos que reconozcan el impacto de los movimientos de precios de los competidores y del valor que el consumidor le da al producto como variables explicativas y que en un corto plazo permite detectar las ventas de las referencias de motos y por consiguiente la participación de mercado de la marca ensambladora.

2.2. Justificación

Si bien los problemas de predicción de demanda no son nuevos en la industria de motos (Rasim et al., 2018), estos se han abordado en su mayoría desde una óptica estática considerando unas condiciones estables del mercado (Gautam et al., 2021) e incluso obviando el dinamismo de la competencia para subir precios o bajarlos por medio de campañas de descuentos puntuales (Di Pillo et al., 2016). Este supuesto de estabilidad ha subvalorado la importancia de valor percibido (relación del beneficio percibido versus precio percibido) que el consumidor le da al producto en cuestión, generando modelos que no permanecen en el tiempo (Marn et al., 2004).

[1] Datos tomados del dataset utilizado para el entrenamiento de los modelos y provista por la empresa a través de su conexión con el RUNT

Construir un modelo de predicción de ventas de motos bajo la óptica del aprendizaje automático que incorpore conceptos de las teorías de precios y del consumidor les permitirá a las compañías de la industria tener una predicción acoplada al dinamismo del mercado (Leszinski & Marn, 2016) y, por tanto, les dará herramientas para diseñar estrategias acertadas de comercialización y planificación.

La importancia de proponer este modelo radica en que la rentabilidad de las decisiones que tomen las compañías respecto al manejo de sus inventarios, procesos productivos y definición de estrategias de precios depende de qué tan acertadas son las predicciones de las ventas (Sharif Azadeh et al., 2015). Si la empresa sobreestima el potencial de ventas de su producto en el mercado incurrirá en costos operativos innecesarios, pero si por el contrario la empresa subestima sus ventas, se puede presentar escasez del producto que lleve al consumidor a buscar un sustituto y por tanto también afectará la rentabilidad esperada (Rasim et al., 2018).

Adicional, el enfoque de Machine Learning permitirá sortear algunos de los inconvenientes que se pueden presentar con el uso de los modelos tradicionales de series de tiempo: Se necesita muchos datos para poder determinar la estacionalidad, se olvidan otros factores externos, se deben limpiar los datos atípicos y outliers (Panarese et al., 2022).

2.3. Objetivos

2.3.1. Objetivo general

Construir un modelo usando técnicas de Machine Learning que permita predecir mensualmente las ventas de motos (vehículos de combustión de 2 ruedas) y participación de mercado de una marca ensambladora del mercado colombiano incorporando variables de la metodología “Customer Value Map”.

2.3.2. Objetivos específicos

- Definir un protocolo de limpieza de datos y tratamiento de outliers teniendo en cuenta el contexto de negocio del mercado de venta de motos.
- Establecer la forma apropiada de incluir el valor percibido por el consumidor (Customer Value Map) como variable explicativa del modelo predictivo y otras características distintas a las unidades vendidas.
- Generar modelos predictivos con distintas técnicas de Machine Learning que permitan predecir las ventas de motos nuevas en Colombia.
- Comparar los modelos construidos por medio de métricas que evalúen el mínimo error en la predicción.

3. Marco teórico

Dado que el objetivo de este proyecto es construir un modelo de predicción de ventas de motos usando técnicas de Machine Learning e incorporando el concepto

“Customer Value Map”, es pertinente aclarar cada uno de los conceptos claves aquí planteados.

En primer lugar, es importante aclarar que el proyecto se ubica en el mercado de motos en Colombia; concretamente el denominado 2W (dos ruedas) de combustión (no se analizarán vehículos eléctricos). Este mercado consta principalmente (más del 90% del mercado) de 7 marcas ensambladoras (AKT, Auteco, Hero, Honda, Suzuki, UMA, Yamaha), las cuales compiten en cada uno de los segmentos. Se considera que una referencia compite con otra si pertenecen al mismo segmento (Mohd et al., 2013).

Los segmentos considerados en el proyecto han sido determinados por el equipo comercial de una de las compañías analizadas:

- Segmento de trabajo: Referencias de motos con cilindraje menor a 150cc y cuyo uso principal es realizar una actividad económica como el transporte de pasajeros y domicilios.
- Segmento cómodo: Referencia de motos con cilindraje entre 150cc y 200cc. Su uso principal está reservado para jóvenes estudiantes como un producto aspiracional para llegar al segmento deportivo.
- Segmento deportivo: Referencias de motos con cilindraje mayor a 200cc y menor a 250cc y cuyo uso principal es ocio y diversión.

Otros conceptos que resultan importantes para enmarcar el proyecto y comprender el objetivo son:

- Forecasting - pronóstico: Se puede definir como la implementación de un proceso que permitirá predecir las salidas futuras de una variable haciendo uso de datos históricos (Buttner & Rabe, 2021).

- Machine Learning: Es una rama de la inteligencia artificial que busca el desarrollo de algoritmos, los cuales utilizan la información pasada para aprender de escenarios futuros. A través de estos algoritmos se definen reglas provenientes de patrones identificados en los datos (Buttner & Rabe, 2021).

Para continuar este estado del arte es necesario hacer una división entre los dos tipos de conceptos bajo los cuales se enmarca el proyecto: En primer lugar, se definirán algunos conceptos relacionados con teorías del consumidor, específicamente con la metodología “Customer Value Map”; en segundo lugar, se introducirán algunas nociones propias de los modelos predictivos, concretamente de modelos de Machine Learning.

3.1. Conceptos de teorías del consumidor:

Si bien las predicciones de ventas de motos podrían hacerse considerando únicamente las variables de ventas rezagadas, la intención de este proyecto es incluir variables propias de las teorías de precios y percepción del consumidor. Lo anterior, teniendo en cuenta que las ventas de un producto están ligadas a la intención de compra de un consumidor, lo cual no solo está asociado al valor monetario de la

referencia en cuestión sino también al proceso de compra que lleva al consumidor a tomar la decisión final (Acquila-Natale & Iglesias-Pradas, 2021).

Así, se puede definir que el valor que un consumidor le da al producto es más que el precio de venta ya que tiene en cuenta las diferencias entre los beneficios y los costos que los consumidores evalúan antes de hacer la compra. En el proceso de compra, la variable dependiente será la intención de compra y las independientes serán los atributos que llevan a cada consumidor a seleccionar un determinado producto (Leszinski & Marn, 2016).

Concretamente, “Customer Value Map” es una metodología de entendimiento del consumidor que permite graficar la intención de compra de un producto sobre otro, dado un valor y un precio percibido por el consumidor (Marn et al., 2004). La representación del “Customer Value Map” se puede evidenciar en la Figura 1, tomada del artículo “Setting value, not Price” (Leszinski & Marn, 2016).

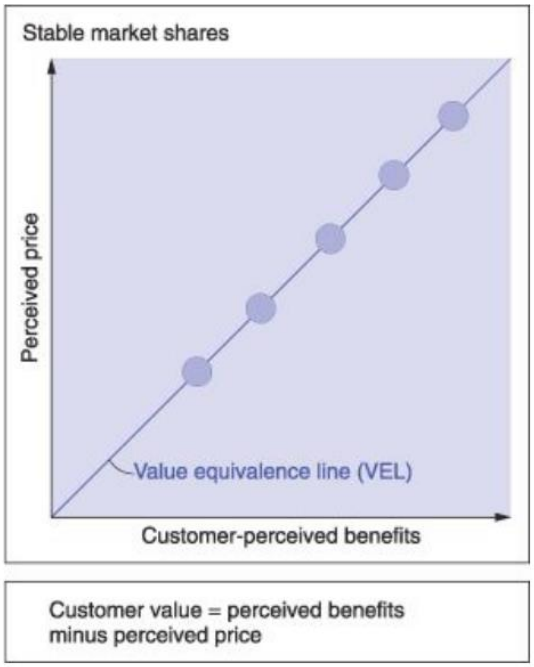

Figura 1: Representación gráfica "Customer Value Map" (Leszinski & Marn, 2016)

En este mapa, el eje X (Customer-perceived benefits) representa los beneficios percibidos por el cliente; estos beneficios son el resultado de la ponderación de atributos que el consumidor considera al momento de la compra. Por otro lado, en el eje Y (Perceived price) se encuentra el precio percibido por el consumidor, que para efectos de este proyecto consiste en el precio del artículo publicado en la página web oficial de la marca ensambladora. Cada punto en el mapa representa cada artículo o referencia que compiten en la decisión de compra del cliente; así, las referencias con mayor precio y valor se encuentran ubicados en la parte superior derecha (Leszinski & Marn, 2016).

Como parte de la teoría “Customer Value Map”, se define que la pendiente de la línea de equilibrio “VEL” (como se define en la Figura 1) la determina la referencia líder del

mercado (la de mayores ventas en el momento dado) (Marn et al., 2004). Esta ubicación en el mapa es la que determina la intención de compra y por tanto será la variable utilizada para complementar el modelo predictivo.

3.2. Conceptos de modelos predictivos

Generalmente, a los problemas de pronósticos de ventas se les ha dado un tratamiento bajo la óptica de modelos de series de tiempo clásicas (Rasim et al., 2018). Algunos autores han explorado otras opciones que van desde puntos de vista cualitativos basados en la experiencia hasta modelos cuantitativos (Navratil & Kolkova, 2019).

Si bien los modelos cualitativos se han utilizado para predecir ventas en algunas compañías, e incluso han complementado los resultados de los modelos cuantitativos, para este proyecto no serán analizados. El interés de este proyecto va orientado hacia los modelos cuantitativos, los cuales se han estudiado desde diferentes enfoques. El enfoque principal y más usado (antes mencionado) es el de series de tiempo. El enfoque de Machine Learning (interés principal de este proyecto), se puede dividir en modelos basados en regresiones (más usados), modelos basados en árboles (Random Forest, Gradient Boosting, Gradient Boosting Machine, LightGBM, XGBoost), modelos de redes neuronales y modelos empíricos construidos bajo la óptica de negocio (Modelo de Uber, Modelo Prophet de Facebook, Modelo Explain) (Navratil & Kolkova, 2019).

Adicional al enfoque, es pertinente analizar el proceso. Muchos autores coinciden en que un protocolo apropiado para la construcción de los modelos de pronósticos de ventas podría ser:

a) Recopilación de datos: Concretamente para la industria de motocicletas se deben recopilar los datos separados para cada uno de los segmentos a los que pertenece la referencia; esto para evitar incluir datos que no aportan poder predictivo al modelo (Rasim et al., 2018).
b) Ingeniería de características: Determinar las características de los datos que deben ser analizados, eliminados o incluidos (Kolková & Navrátil, 2021).
c) Aplicación de las técnicas de predicción.
d) Evaluación del resultado del pronóstico mediante el cálculo de las métricas de precisión (Rasim et al., 2018).

La ingeniería de características para los modelos de predicción de ventas de motos ha tenido especial cuidado en el tratamiento de los outliers, tratando de analizar si la existencia de este es una característica más del modelo. Como parte de la ingeniería de características, se proponen también métodos de agrupamiento automático que permitan identificar patrones. Adicional, se intentan capturar características de las ventas en días, semanas, meses y años (Dairu & Shilong, 2021). Estos outliers se detectan bajo dos modalidades: el conocimiento del mercado de motos en el contexto temporal estudiado (Sharif Azadeh et al., 2015) y un análisis de los datos mismos usando método de variación y desviación estándar (Dairu & Shilong, 2021).

Algunos autores consideran los efectos de los bajos niveles de inventario de una marca como un outlier que debe analizarse, ya que, bajo este fenómeno, la demanda real puede diferir de las ventas registradas; los consumidores buscarán un sustituto del producto o se abstendrán de realizar la compra. Para este caso, el dato atípico no es eliminado del modelo, sino que se incorpora al mismo con técnicas que expliquen el fenómeno desde una descomposición estacional (Sharif Azadeh et al., 2015).

Otros autores tratan este fenómeno de inventario como un problema de disponibilidad del producto y hacen una imputación de datos para las ventas del producto con baja disponibilidad ajustando los valores con las variaciones en las ventas del producto sustituto en el periodo de tiempo en el que se presentó el caso (Gautam et al., 2021).

También se encontraron autores que optan por una solución más radical. Deciden eliminar los datos atípicos haciendo pruebas de robustez al modelo que garanticen su eficiencia (Kolková & Navrátil, 2021).

Por último, hay autores que identifican los periodos de promociones y descuentos como un fenómeno anómalo que generará datos atípicos en los datos. Si estos eventos se repiten recurrentemente se pueden incluir en el modelo como una característica adicional de los datos que explican la variable de respuesta (Pavlyshenko, 2019).

Siguiendo con el proceso planteado para la construcción de modelos predictivos, es preciso definir algunos conceptos relacionados con las **técnicas de Machine Learning** exploradas. Como se planteó anteriormente, estas técnicas se agrupan en diferentes enfoques. Los enfoques de *regresión* suponen el comportamiento lineal del fenómeno estudiado. Bajo esta óptica se encuentran las regresiones lineales tradicionales que incluyen como variables explicativas las ventas de las motos (incluso con rezagos) y que van muy ligadas al enfoque de series de tiempo clásicas. Así mismo, se encuentran las regresiones con algún tipo de penalización con el fin de garantizar la precisión del modelo (Mohd et al., 2013). Los *enfoques de árboles*, en los que se diseña un algoritmo con forma de árbol con la intención de predecir o clasificar (Panarese et al., 2022). Para este caso se exploran las siguientes técnicas:

- Random Forest (RF): Es una técnica que construye múltiples árboles aleatorios, cuyas predicciones se agregan para reducir la varianza (Buttner & Rabe, 2021). La predicción final se obtiene de la votación de los árboles individuales. Es una técnica comúnmente usada dada su facilidad de entrenar y a su capacidad de evitar sobre ajuste (Pavlyshenko, 2019).
- Gradient Boosting (GB): Se basa en el cálculo de los residuos como diferencia entre el valor real y el valor aproximado, donde el objetivo del algoritmo es disminuir la función de pérdida usando aproximaciones refinadas de forma iterativa. El proceso de entrenamiento se realiza iterativamente, con cada árbol de decisión ajustado para mejorar el error del modelo anterior (Panarese et al., 2022).
- Gradient Boosting Machine (GBM): Es una técnica que construye y promedia varios modelos, mejorando la función objetivo en cada iteración. es una variante de GB que se enfoca en ajustar un modelo de árbol de decisión a los datos (Kiely & Bastian, 2019).

- LightGBM: Utiliza una técnica de división en hojas para reducir la complejidad y el tiempo de entrenamiento (Jiang et al., 2021). Particularmente (Deng et al., 2021) introduce el concepto de predicción de ventas bajo esta técnica incluyendo las ventas rezagadas como una de las características del modelo.
- XGBoost: Es un árbol de decisión que cuenta con las mismas reglas que el clásico árbol de decisión. El árbol de cada nodo interno representa los valores para la prueba y el nodo hoja representa la decisión (con puntajes). Esta técnica hace uso de regresiones lineales para aproximar los residuos de los árboles (Dairu & Shilong, 2021).

En cuanto a los enfoques en *redes neuronales* las exploradas son: Red neuronal convolucional - CNN (Menculini et al., 2021); LSTM que es un tipo de red neuronal recurrente con aplicaciones en series de tiempo ya que adopta las no linealidades de la serie (Hern Kong et al., 2021); y los modelos NNTEAR planteados por (Kolková & Ključnikov, 2022), los cuales a través del uso de redes neuronales permiten establecer relaciones no lineales complejas entre una variable y sus predictores. Los modelos NNTEAR consideran solo redes neuronales de alimentación directa con una capa oculta, utilizando los valores rezagados de la serie como entradas de la red. El modelo se define como Nnetar (p, k), con p entradas rezagadas y k nodos en la capa oculta (Kolková & Ključnikov, 2022).

Por último, se encuentran las técnicas con enfoque en *modelos empíricos*:

- El modelo de Uber: Utiliza el sistema de grafos de una red neuronal dinámica, combinando el modelo de suavización exponencial con redes neuronales avanzadas de memoria de corto plazo (Kolková & Ključnikov, 2022).
- El modelo de Facebook (prophet): Se utiliza para predecir valores de una serie de tiempo con tendencia y estacionalidad. Se construye a través de la combinación no lineal de componente tales como la tendencia, la estacionalidad y las vacaciones (Navratil & Kolkova, 2019).
- Metodología Explain: Expuesta por (Bohanec et al., 2017). La técnica clave es el análisis de sensibilidad orientado al cambio de los inputs del modelo con el fin de observar los cambios en la salida.

Una vez exploradas las técnicas aplicables al problema, es necesario definir las **métricas** bajo las cuales se evaluarán los modelos construidos. Al respecto se encuentra que estos métodos se pueden dividir en medidas dependientes de la escala, basados en errores porcentuales o escalados y basados en errores relativos (Navratil & Kolkova, 2019).

Las métricas analizadas son:

- Error medio (ME): Es la diferencia entre la predicción y el valor real promedio. Puede arrojar resultados negativos en caso de que la predicción sea mayor que el valor real (Kolková & Navrátil, 2021).
- Error cuadrático medio (MSE): Es la media de los errores cuadrados entre la predicción y el valor real. Se calcula sumando los errores al cuadrado y dividiéndolos por el número de muestras (Kolková & Navrátil, 2021).

- Error absoluto medio (MAE): es la media de los errores absolutos entre la predicción y el valor real. Se calcula sumando los errores absolutos y dividiéndolos por el número de muestras (Kolková & Navrátil, 2021).
- Error porcentual medio (MPE): Es la diferencia porcentual entre la predicción y el valor real promedio (Kolková & Navrátil, 2021).
- Error porcentual absoluto medio (MAPE): Se calcula sumando los errores porcentuales absolutos y dividiéndolos por el número de muestras (Rasim et al., 2018).
- Raíz del error cuadrático medio (RMSE): Es la raíz cuadrada del MSE. Se utiliza para obtener una medida de error más fácil de interpretar en las mismas unidades que la variable pronosticada (Menculini et al., 2021).

4. Metodología

En el desarrollo de este proyecto se hará uso de la metodología CRISP-DM (IBM, 2020). La cual consta de 6 pasos secuenciales que se describen a continuación:

a) **Comprensión del negocio**: En esta primera fase se describe el problema de negocio que se intenta resolver y los objetivos que le apuntan a la solución de este. Como se ha planteado anteriormente, el problema de negocio es poder realizar una predicción de las ventas de motos para el mercado colombiano incorporando teorías de conocimiento del consumidor ("Consumer Value Map") y técnicas de Machine Learning.
b) **Comprensión de los datos:** En esta fase se hace un análisis de los datos disponibles y se hace una recolección de estos. Se evalúa inicialmente la calidad del *dataset* construido y se corre un análisis descriptivo de los datos con el objetivo de hacer un análisis exploratorio que permita identificar patrones. Se identifica que existe un periodo particular en la serie de tiempo, comprendido por el periodo de pandemia en el que las ventas disminuyeron considerablemente para todas las marcas competidoras.
c) **Preparación de los datos:** En esta fase se limpian los datos y se hacen las transformaciones necesarias para facilitar el modelado. Para el caso del *dataset* requerido, se ve la necesidad de aplicar técnicas de homologaciones de referencias y agrupamientos de *dataset* dado que los datos provienen de diferentes fuentes.
d) **Modelado:** En esta fase se modelarán las técnicas de Machine Learning basadas en árboles exploradas en el marco teórico.
e) **Evaluación:** En esta fase se calcularán las métricas de precisión exploradas en el marco teórico con el objetivo de comparar los resultados de los modelos construidos en la fase anterior.
f) **Despliegue:** Si bien el objetivo de este proyecto llega hasta la construcción y evaluación del modelo, para esta fase se compartirán los resultados con los equipos de analítica y comerciales de la organización con el fin de diseñar la mejor manera de implementarlo.

5. Experimentación

5.1. Datos y definiciones de negocio

Los datos fueron proporcionados al estudiante por medio de los accesos que este cuenta con la empresa a la que está vinculado (respectiva autorización de la compañía). Estos datos refieren principalmente a las unidades vendidas en el mercado de motos nuevas en Colombia y a algunas características del consumidor que la empresa obtuvo a través de un proyecto de investigación de mercado.

Los datos utilizados en este estudio se encuentran distribuidos en 3 conjuntos de datos con formato CSV:

- ***Dataset 1***: Ventas históricas diarias de cada una de las referencias de las marcas estudiadas. Datos desde enero 2019 hasta abril 2023.
- ***Dataset 2*****:** Precios de lista y promociones de cada una de las referencias estudiadas. Datos desde enero 2019 hasta abril 2023.
- ***Dataset 3*****:** Valor percibido para cada una de las referencias estudiadas.

Dado que por confidencialidad de la información no se permite publicar el nombre de la empresa dueña de los datos y objeto de este análisis, el primer proceso realizado al conjunto de datos proporcionados fue el de anonimizar los datos:

- Los nombres de las empresas estudiadas fueron reemplazados por EMPRESA1, EMPRESA2, EMPRESA3 y así sucesivamente.
- Los nombres de las marcas estudiadas fueron reemplazados por MARCA1, MARCA2, MARCA3 y así sucesivamente.
- Los nombres de las referencias de motos se encuentran como REF1, REF2, REF3 y así sucesivamente para cada línea.

Las variables de las cuales se dispone en el *Dataset 1* son:

- Fecha de venta: Corresponde a la fecha en la cual se realizó la venta de la moto al consumidor final. Formato ‘dd-MM-yyyy’.
- Categoría: Indica si el artículo es una motocicleta, motocarro o patineta.
- Línea RUNT: Nombre de la referencia de moto según los registros del RUNT.
- Empresa: Nombre de la empresa que ensambló la referencia.
- Marca: Corresponde a la marca de la referencia.
- Segmento RUNT: Es el segmento por el cual el RUNT categórica la referencia.
- Cantidad: Es la suma de las ventas de la referencia en esa fecha.

Las variables de las cuales se dispone en el *Dataset 2* son:

- Fecha: Corresponde a la fecha (diaria) en la cual se consultó el precio de la referencia. En formato ‘dd-MM-yyyy’.

- Línea: Nombre de la referencia de moto según la información publicada en las diferentes páginas web de las marcas analizadas.
- Precio en lista: Corresponde al precio del día en pesos colombianos de la referencia.
- Bono: Corresponde al descuento (en pesos colombianos) del día para cada una de las referencias.

Las variables de las cuales se dispone en el *Dataset 3* son:

- Fecha de la investigación de mercado: Fecha en la cual se realizó el proyecto de investigación de mercado y por tanto se obtuvo el dato.
- Línea: Nombre de la referencia de moto según la información publicada en las diferentes páginas web de las marcas analizadas.
- Valor percibido: Número (entre 0 y 500) que indica el valor que los consumidores le dan al producto (referencia de moto). Entre más grande el número hay un mayor valor.
- Segmento: Corresponde a la segmentación del mercado a la que pertenece cada referencia. Este dato fue construido por la empresa (diferente al que se encuentra en las bases del RUNT) luego de un proceso de entendimiento del mercado.

Por último, se vio la necesidad de construir un *Dataset 4* para poder unificar los datos. Este último *dataset* se construyó durante el desarrollo del proyecto con ayuda del equipo comercial de la empresa:

- Línea RUNT: Nombre de la referencia de moto según los registros del RUNT.
- Línea: Nombre de la referencia de moto según la información publicada en las diferentes páginas web de las marcas analizadas.

De esta manera se logró construir una llave única que permite unir los 3 conjuntos de datos anteriormente descritos.

5.2. Preparación de los datos e Ingeniería de características

Inicialmente se realiza un proceso de unificación de los 3 conjuntos de datos haciendo uso del *Dataset* 4 con las homologaciones. En la Figura 2 se ilustra el proceso en el cual se une el *dataset* 1 y el *dataset* 2 con la fecha de venta de la moto y se llega a un *dataset* 4 que unifica la variable “Línea RUNT” del *dataset* 1 con la variable “Línea” del *dataset* 2 para cada una de las fechas:

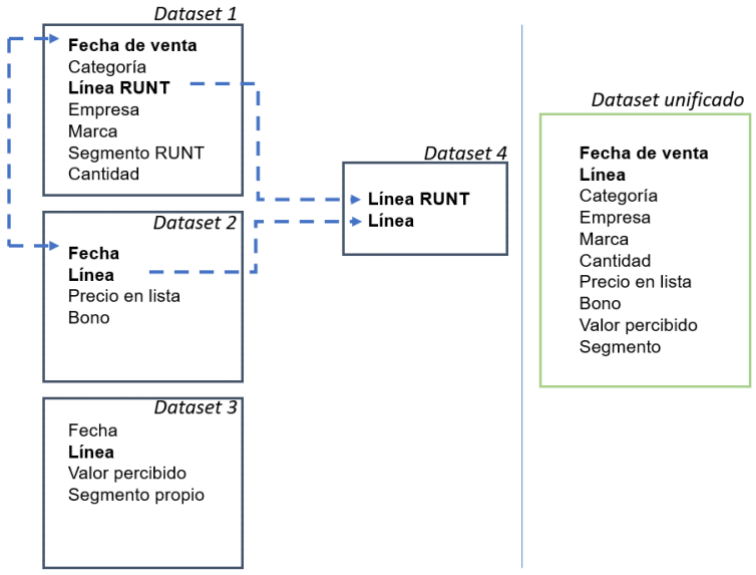

Figura 2: Unificación de los datos

Adicional, se hace un proceso para transformar la fecha diaria en semanas del año, de esa manera se trabaja con un conjunto de datos con periodicidad semanal. En la Figura 3 se observa una previsualización del *dataset* luego de la unificación:

	Año	Semana	LINEA	CATEGORIA	EMPRESA	MARCA	SEGMENTO	SUBCATEGORIA	Cantidad	Precio_Real	Valor percibido	Segmento Mapas
2366	2023	4	REF12	MOTORCYCLES	MARCA2	MARCA2	COMMUTER ENTRY	STREET	187	5720000.00	405.37	COMMUTER ENTRY
2368	2023	4	REF14	MOTORCYCLES	EMPRESA3	MARCA9	TRABAJO	STREET	304	5999000.00	475.79	COMMUTER ENTRY
2374	2023	4	REF19	MOTORCYCLES	MARCA6	MARCA6	TRABAJO	STREET	140	NaN	376.95	COMMUTER ENTRY
2386	2023	4	REF30	MOTORCYCLES	EMPRESA2	MARCA10	TRABAJO	STREET	63	4999000.00	390.74	COMMUTER ENTRY
2390	2023	4	REF33	MOTORCYCLES	EMPRESA1	MARCA8	TRABAJO	STREET	69	5899999.00	429.47	COMMUTER ENTRY

Figura 3: Vista previa de los datos

El proceso previo al procesamiento de los datos se observa en la Tabla 1:

PROCESO	TRATAMIENTO APLICADO
PERIODO DE ANÁLISIS	Se decide trabajar con datos desde enero del año 2022, esto debido a que en años anteriores el mercado de motos (principalmente para la empresa que proporcionó los datos) tuvo un cambio importante en su estructura (algunas marcas importantes dejaron de ser ensambladas por la compañía y se acoplaron otras).
VACÍOS EN LA COLUMNA "VALOR PERCIBIDO"	Si esta columna tiene vacíos implica que no se hizo la investigación de mercado para esa referencia, por tanto, se eliminan estos datos.

PROCESO	TRATAMIENTO APLICADO
COLUMNA "PRECIO_REAL"	El precio real de la referencia es la resta entre el precio en lista y el bono (Descuento) en la fecha de la venta.
LIMPIEZA DE DATOS ADICIONAL	Algunas combinaciones de empresa y referencia no tenían sentido para el negocio y por tanto fueron eliminadas Precios inferiores a $500 fueron reemplazados por el precio de la semana anterior.
DATOS VACÍOS	Se imputó valores para precios vacíos con el último precio de la referencia de moto.
VACÍOS EN LA COLUMNA "SEGMENTO MAPAS"	Se eliminaron del estudio dado que no fueron tenidos en cuenta en la investigación de mercado para conocer la percepción del consumidor.

Tabla 1: Preprocesamiento de los datos

Posterior a este procesamiento de los datos se decide incluir las variables del "Customer Value Map", de tal manera que conserven las características de las teorías expuestas por (Marn et al., 2004) y (Leszinski & Marn, 2016). A continuación, se recopilan las definiciones del "Customer Value Map" y la manera en que fueron acopladas en el análisis:

1. El mercado se analiza a través de la participación porcentual de las unidades vendidas de cada referencia sobre el total del segmento (Market Share) y no a través de las unidades absolutas vendidas en el periodo de tiempo seleccionado (Marn et al., 2004): Se transforma la columna "Cantidad" en la columna Market Share (porcentaje del mercado de cada referencia en cada semana) usando la Ecuación 1 para cada referencia en el segmento.

$$Market\ Share = \frac{Cantidades\ de\ la\ referencia}{Total\ cantidades\ del\ segmento}$$

Ecuación 1: Definición de Market Share

2. El líder del mercado es la referencia que vende la mayor cantidad de unidades en el segmento (Leszinski & Marn, 2016): Se calculan los líderes semanales para cada segmento, identificando aquellos que tienen un mayor porcentaje de ventas en el periodo (Market Share).
3. La línea de referencia (VEL) es una línea recta cuyo intercepto es la coordenada (0,0) y su pendiente (*m*) está dada por el líder del mercado (Marn et al., 2004): se calcula la pendiente por medio de la Ecuación 2.

$$m = \frac{Precio\ del\ líder\ en\ la\ semana}{Valor\ percibido\ del\ líder\ en\ la\ semana}$$

Ecuación 2: Pendiente de la línea VEL

4. La posición de la referencia respecto a la línea de referencia (VEL) determinará si esa referencia perderá o ganará participación de mercado. Referencias que se encuentren a la derecha de VEL ganarán Market Share, si se encuentran a la izquierda perderán, si se encuentra sobre la línea es el líder (Marn et al., 2004): Se construye la variable posición respecto a las coordenadas de cada referencia en el "Customer Value Map"; 0) posición derecha, 1) posición izquierda, 2) es el líder.
5. En el "Customer Value Map", entre más alejada (distancia a la recta) esté la coordenada de la referencia a la línea VEL, mayor será la propensión para perder Market Share (Marn et al., 2004): Se construye la variable distancia como la distancia perpendicular (euclídea) del punto (referencia de la moto) a la línea VEL. Se usa la Ecuación 3 para cada una de las referencias en el conjunto de datos.

$$Distancia = \frac{|m * Valor\ percibido - Precio\ real|}{\sqrt{m^2 + 1}}$$

Ecuación 3: Distancia perpendicular de un punto a la recta VEL

6. El "Customer Value Map" es una herramienta dinámica que no predice el comportamiento del mercado en el instante, sino que determina que en el futuro y de acuerdo con la posición y distancia de la referencia a la línea VEL se ganará o se perderá Market Share (Leszinski & Marn, 2016): Se crean nuevas columnas con los rezagos semanales (hasta 4 rezagos) de las variables "Precio Real" y "Distancia". Dado que esta transformación recorta la cantidad de datos, se procede a eliminar los vacíos.

En este punto, se define:

- **Variable explicada:** Market Share (Cantidades vendidas representadas como un porcentaje del segmento)

- **Variables explicativas:** conjunto de variables de la teoría de precios ("Customer Value Map") y propias del mercado de motos. Distancia (dist_VEL), Precio_Real, Valor percibido, Precio con 1 rezago (Precio_1), Precio con 2 rezagos (Precio_2), Precio con 3 rezagos (Precio_3), Precio con 4 rezagos (Precio_4), Distancia con 1 rezago (dist_1), Distancia con 2 rezagos (dist_2), Distancia con 3 rezagos (dist_3), Distancia con 4 rezagos (dist_4), Posición.

Luego de incorporar las nuevas variables al modelo, se calculan las correlaciones de estas con la variable explicada (excepto la variable posición que es una codificación de 3 categorías). La matriz de correlación (Spearman) se muestra en la Figura 4:

	Market_Share	dist_VEL	Precio_Real	Valor percibido	Precio_1	Precio_2	Precio_3	Precio_4	dist_1	dist_2	dist_3	dist_4
Market_Share	1.00	-0.32	0.14	0.32	0.13	0.12	0.11	0.12	-0.32	-0.32	-0.32	-0.31
dist_VEL	-0.32	1.00	0.27	-0.65	0.25	0.24	0.23	0.22	0.93	0.89	0.87	0.85
Precio_Real	0.14	0.27	1.00	0.22	0.96	0.93	0.92	0.92	0.23	0.21	0.19	0.19
Valor percibido	0.32	-0.65	0.22	1.00	0.21	0.21	0.20	0.19	-0.66	-0.66	-0.67	-0.66
Precio_1	0.13	0.25	0.96	0.21	1.00	0.96	0.93	0.92	0.26	0.23	0.20	0.20
Precio_2	0.12	0.24	0.93	0.21	0.96	1.00	0.96	0.93	0.25	0.25	0.22	0.21
Precio_3	0.11	0.23	0.92	0.20	0.93	0.96	1.00	0.96	0.23	0.24	0.25	0.23
Precio_4	0.12	0.22	0.92	0.19	0.92	0.93	0.96	1.00	0.22	0.22	0.23	0.25
dist_1	-0.32	0.93	0.23	-0.66	0.26	0.25	0.23	0.22	1.00	0.93	0.89	0.85
dist_2	-0.32	0.89	0.21	-0.66	0.23	0.25	0.24	0.22	0.93	1.00	0.93	0.87
dist_3	-0.32	0.87	0.19	-0.67	0.20	0.22	0.25	0.23	0.89	0.93	1.00	0.91
dist_4	-0.31	0.85	0.19	-0.66	0.20	0.21	0.23	0.25	0.85	0.87	0.91	1.00

Figura 4: Correlación de las variables explicativas con la variable explicada

Se hace una selección de variables basados en las correlaciones y en las recomendaciones de los expertos del mercado en la empresa analizada. Las variables seleccionadas son:

- Posición
- Precio con 1 rezago (Precio_1)
- Precio con 2 rezago (Precio_2)
- Distancia con 1 rezago (dist_1)

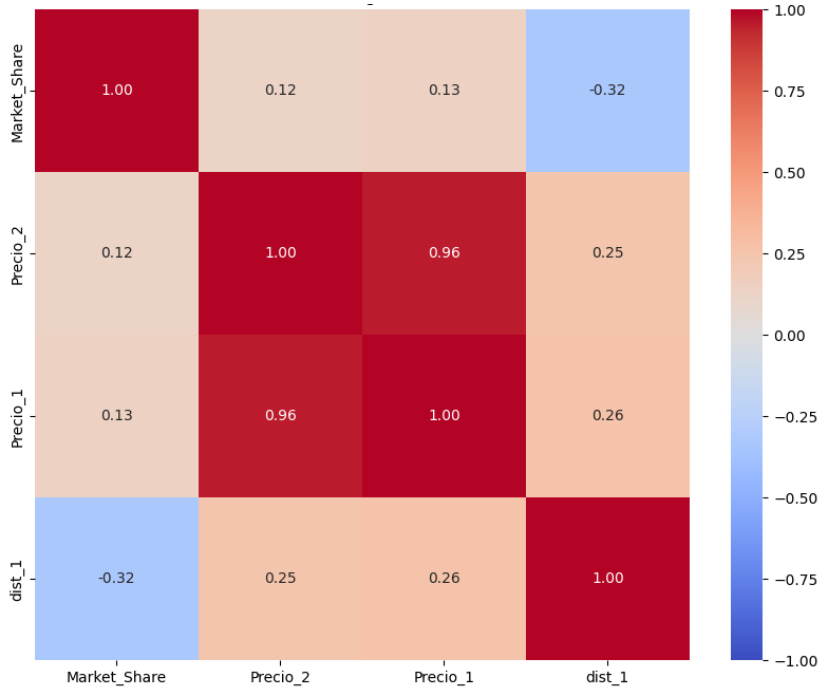

Figura 5: Correlación de las variables seleccionadas con la variable explicada

Con el correlograma de la Figura 5 se puede observar que una de las teorías del "Customer Value Map" descritas por (Marn et al., 2004) es cierta. Existe una correlación negativa entre la distancia y el Market share (la propensión a ganar participación de mercado), es decir, entre mayor sea la distancia de la referencia de moto a la línea VEL, mayor será la propensión para perder Market Share.

5.3. Modelado

La serie de tiempo con las ventas expresadas en Market Share se observa en la Figura 6:

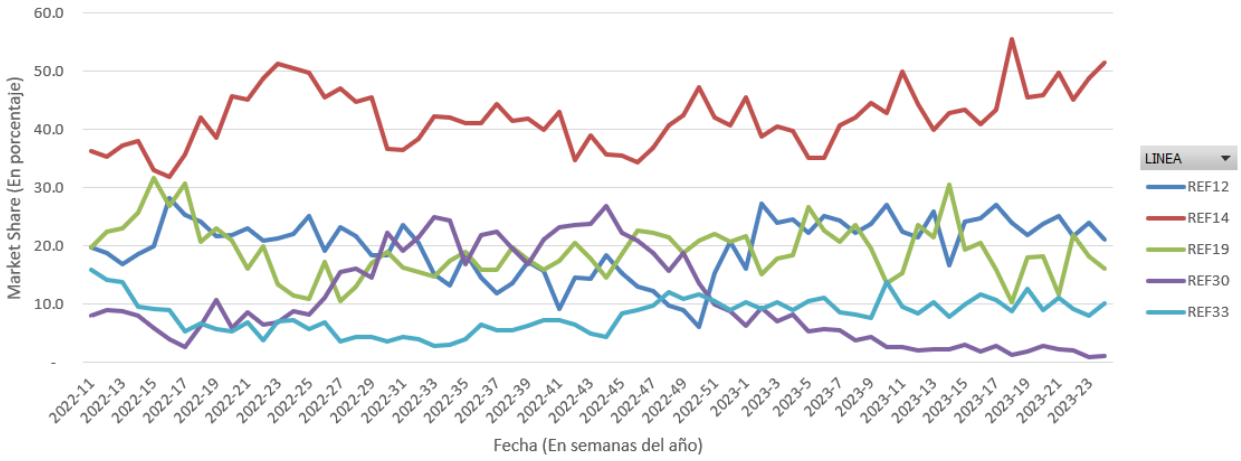

Figura 6: Serie de tiempo con las ventas semanales expresada en Market Share

Así mismo, la descomposición de cada una de las series (para cada referencia) se observa a continuación (Figuras de la 7 a la 11), donde se observa que en general para el 2023 hubo un aumento en las ventas de las referencias, exceptuando la REF19 y REF30:

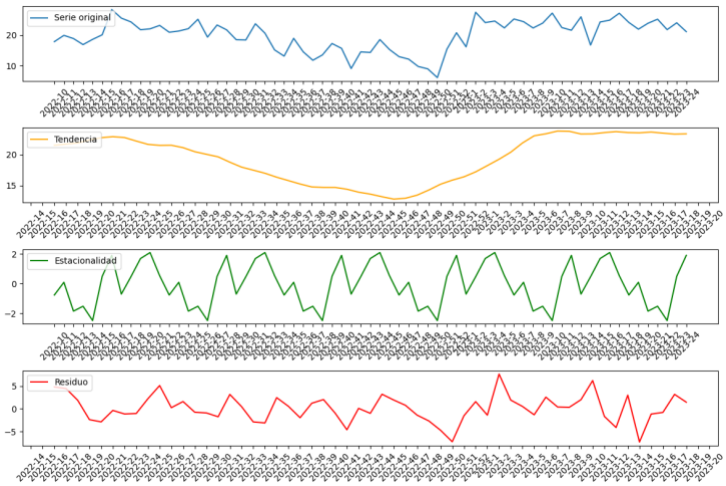

Figura 7: Descomposición de la serie REF12

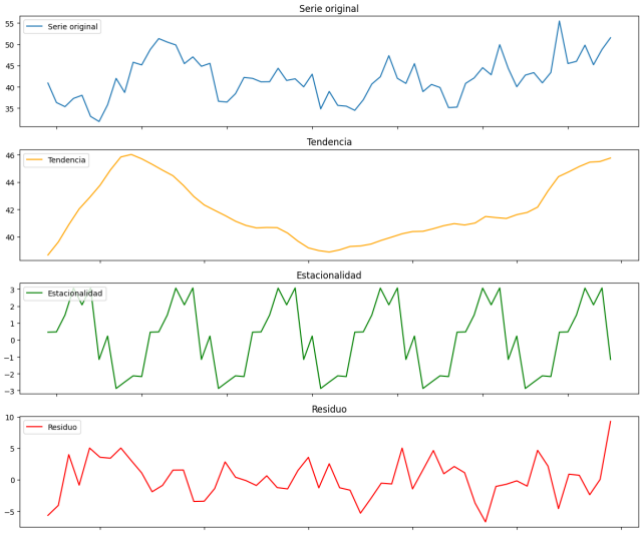

Figura 8: Descomposición de la serie REF14

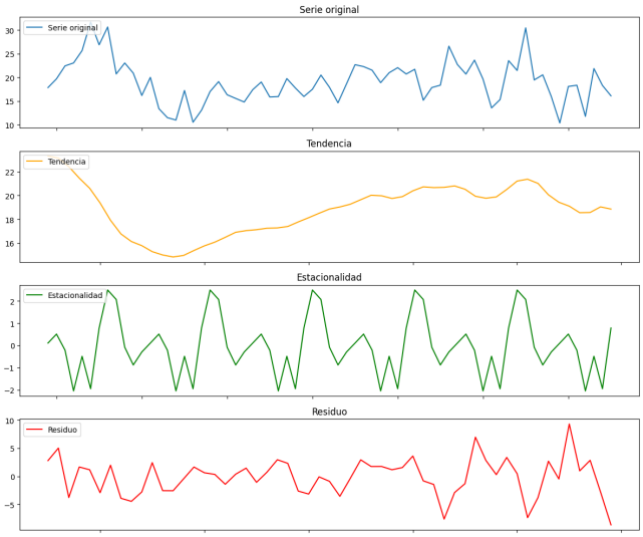

Figura 9: Descomposición de la serie REF19

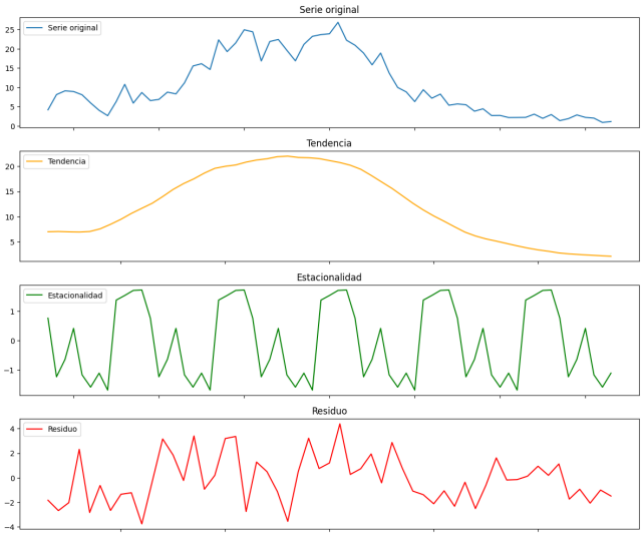

Figura 10: Descomposición de la serie REF30

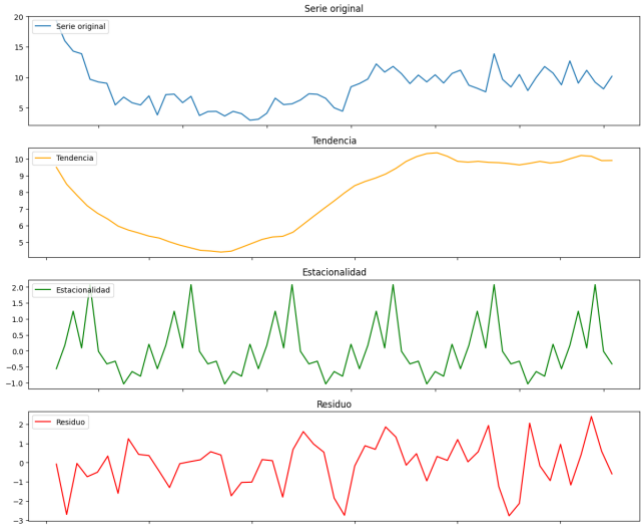

Figura 11: Descomposición de la serie REF33

Se identifica un líder absoluto del segmento. La REF14 es el líder en ventas para cada una de las semanas analizadas. Adicionalmente, se observa que la REF30 pasa por diferentes momentos de ventas, inicialmente con un Market Share bajo hasta el periodo 2022-31 en el cual se convierte en el segundo en ventas, pero en el periodo 2022-51 queda de último en ventas.

Con esta serie y con las variables seleccionadas se procede a hacer la modelación del problema bajo técnicas de Machine Learning basadas en árboles. Inicialmente, se separa el conjunto de datos en entrenamiento y pruebas. Para el conjunto de pruebas se reservan los datos de las fechas '2023-17', '2023-18', '2023-19', '2023-20', '2023-21', '2023-22', '2023-23', '2023-24'. Esta estrategia de segmentación del set de pruebas y entrenamiento se hizo garantizando tener al menos el 10% de los datos de cada referencia para la fase de pruebas y verificando con los expertos comerciales que no hubiera alguna situación especial de mercado en este periodo de tiempo (problemas de abastecimiento, nuevas leyes, incursión de nuevas referencias de motos, entre otras).

La estrategia de modelación se puede observar en la Figura 12:

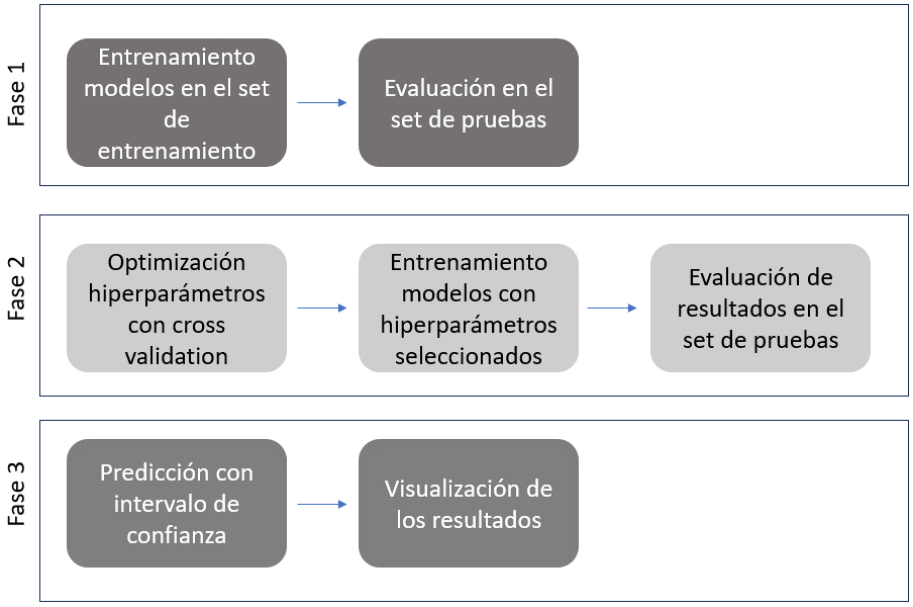

Figura 12: Estrategia de modelación

Como parte de la estrategia de modelación también se decidió utilizar un random_state igual para todos los modelos.

Las técnicas de Machine Learning basadas en árboles que se usaron en la estrategia de modelación son:

- Random Forest
- Gradient Boosting
- LightGBM
- XGBoost

Adicional a estas, se entrenó como modelo inicial una regresión lineal con el objetivo de comparar los resultados. La estrategia usada en cada fase (Figura 12) se describe a continuación:

Fase 1:

En esta primera Fase se entrenaron los modelos de Machine Learning con hiperparámetros iniciales seleccionados por criterio del estudiante (ver Tabla 2) :

	RANDOM STATE	N ESTIMATORS	MAX DEPTH	LEARNING RATE	NUM LEAVES	CRITERION
RANDOM FOREST	59	100	3			gini
GRADIENT BOOSTING	59	100	3	0.1		
LIGHTGBM	59	100		0.05	31	
XGBOOST	59	100	3	0.1		

Tabla 2: Hiperparámetros iniciales

Fase 2:

Para la segunda fase de la estrategia de implementación se busca hacer optimización de hiperparámetros usando la técnica RandomizedSearchCV de sklearn.model_selection con 100 iteraciones. Para todos los modelos explorados se definió la siguiente configuración de RandomizedSearchCV:

- n_iter: Se define 100 iteraciones como el número de configuraciones de hiperparámetros que se deben muestrear.
- Scoring: Se define neg_mean_absolute_error (versión negativa del MAE) como la métrica que se utilizará para evaluar el rendimiento de cada combinación de hiperparámetros y que por tanto se buscará maximizar.
- CV: para la estrategia de validación cruzada se usa el iterador GroupTimeSeriesSplit, esto debido a que es necesario que en las divisiones de la serie se considere la referencia de la moto (variable LINEA).
- n_jobs: Se define en 1 como el número de núcleos de CPU que se utilizarán para realizar la búsqueda en paralelo, es decir que la búsqueda de hiperparámetros se hará de manera secuencial y no en paralelo.

Las configuraciones particulares de cada modelo en el RandomizedSearchCV se encuentran en la Tabla 3 (*randint* se usa para definir un rango de número enteros) :

	RANDOM STATE	N ESTIMATORS	MAX DEPTH	LEARNING RATE	NUM LEAVES	CRITERION
RANDOM FOREST	59	randint(10, 100)	randint(10, 30)			gini
GRADIENT BOOSTING	59	randint(10, 100)	randint(2, 30)	[0.001, 0.01, 0.1, 0.5, 1.0]		
LIGHTGBM	59	randint(10, 100)	randint(2, 30)	[0.001, 0.05 ,0.01, 0.1, 0.5, 0.8, 1.0]	randint(10, 40)	
XGBOOST	59	randint(10, 100)	randint(2, 30)	[0.001, [0.001, 0.05 ,0.01, 0.1, 0.5, 1.0]		

Tabla 3: Configuración de hiperparámetros para RandomizedSearchCV

Adicional a lo descrito en la Tabla 3, se utilizó un “subsample” y ” colsample_bytree” en XGBoost de [0.5, 0.7, 0.8, 1.0].

Finalmente, los hiperparámetros con los que se entrenaron los modelos en la fase de 2 de la estrategia de modelación se encuentran en la Tabla 4:

	RANDOM STATE	N ESTIMATORS	MAX DEPTH	LEARNING RATE	NUM LEAVES	CRITERION
RANDOM FOREST	59	53	12			gini
GRADIENT BOOSTING	59	53	10	0.1		
LIGHTGBM	59	18	26	0.8	17	
XGBOOST	59	45	11	0.1		

Tabla 4: Hiperparámetros seleccionados para cada modelo

Fase 3:

En la fase 3 de la modelación se buscó complementar las predicciones de la fase 2 con la construcción de intervalos de predicción. Para esto, se hizo uso del módulo MapieRegressor de la librería MAPIE. Los hiperparámetros configurados para esta fase en cada uno de los modelos se observan en la Tabla 5:

	Hiperparámetro
estimator	Modelos previamente entrenados (fase 2): Random Forest, Gradient Boosting, LightGbm, XGboost
alpha	0.05
method	Base
cv	Método de remuestreo BlockBootstrap: n_resamplings = 10 n_blocks = 10 overlapping = False (los bloques no se superpondrán) random_state = 59

Tabla 5: Hiperparámetros utilizados en la definición de intervalos de predicción

5.4. Análisis de resultados

Los resultados obtenidos con la implementación de la estrategia de modelación de la Figura 12 para la fase 1 y 2 se pueden observar en la Tabla 6 :

	FASE 1		**FASE 2**	
	R^2	**MAE**	**R^2**	**MAE**
RANDOM FOREST	0.78	6.31	0.91	3.46
GRADIENT BOOSTING	0.92	3.45	0.93	2.94
LIGHTGBM	0.88	4.55	0.91	3.70

XGBOOST	0.91	3.93	0.92	3.14

Tabla 6: Comparación de resultados para la fase 1 y 2 de implementación de modelación

La fase 2 con optimización de hiperparámetros usando RandomSearchCV presenta mejores resultados que la Fase 1 (hiperparámetros iniciales elegidos por intuición) en términos de MAE y de R^2, por lo tanto, se decide elegir estos hiperparámetros para continuar con la fase 3. La evaluación completa de los modelos con la fase 2 se observa en la Tabla 7:

	R^2	MSE	MAPE	RMSE	MAE
LINEAR REGRESSION	0.66	87.91	0.95	9.38	7.41
RANDOM FOREST	0.91	22.52	0.43	4.75	3.46
GRADIENT BOOSTING	0.93	18.63	0.27	4.32	2.94
LIGHTGBM	0.91	22.28	0.78	4.72	3.70
XGBOOST	0.92	20.37	0.34	4.51	3.14

Tabla 7: Evaluación de modelos en el set de pruebas

La evaluación de los modelos basados en árboles es muy similar en todas las métricas y supera en resultados al modelo de regresión lineal. Para el interés particular del problema, el MAE es la medida que se quiere minimizar dado que muestra el error en términos de puntos del Market Share. El modelo con menor MAE es el Gradient Boosting, aunque todos los demás tienen resultados muy similares. En las Figura 13, Figura 14 y Figura 15 se presentan los resultados de la serie para los modelos de Gradient Boosting y XGBoost (modelos con menor MAE):

Figura 13: Predicción para REF12 Y REF14

Figura 14: Predicción para REF19 y REF30

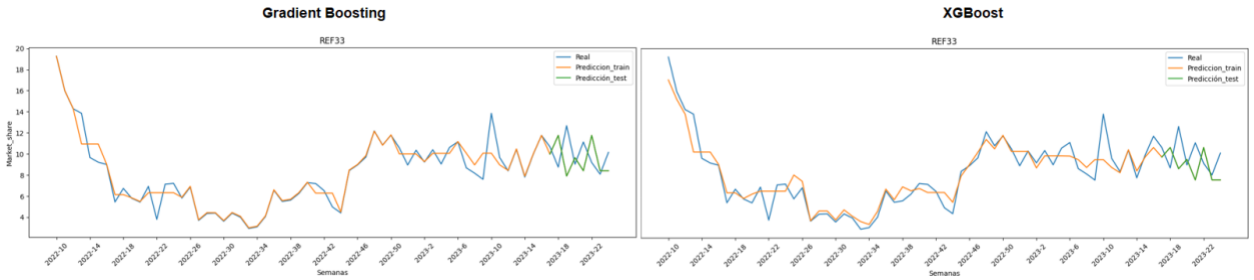

Figura 15: Predicción para REF33

Nuevamente se evidencia que los resultados son muy similares entre los dos modelos, sin embargo, los resultados de la serie para la REF14 y REF19 son los más alejados de la serie original. Con la tercera fase de la estrategia de modelación (intervalos de predicción) se puede mejorar los resultados. Estas series se observan en las Figura 16, Figura 17 y Figura 18:

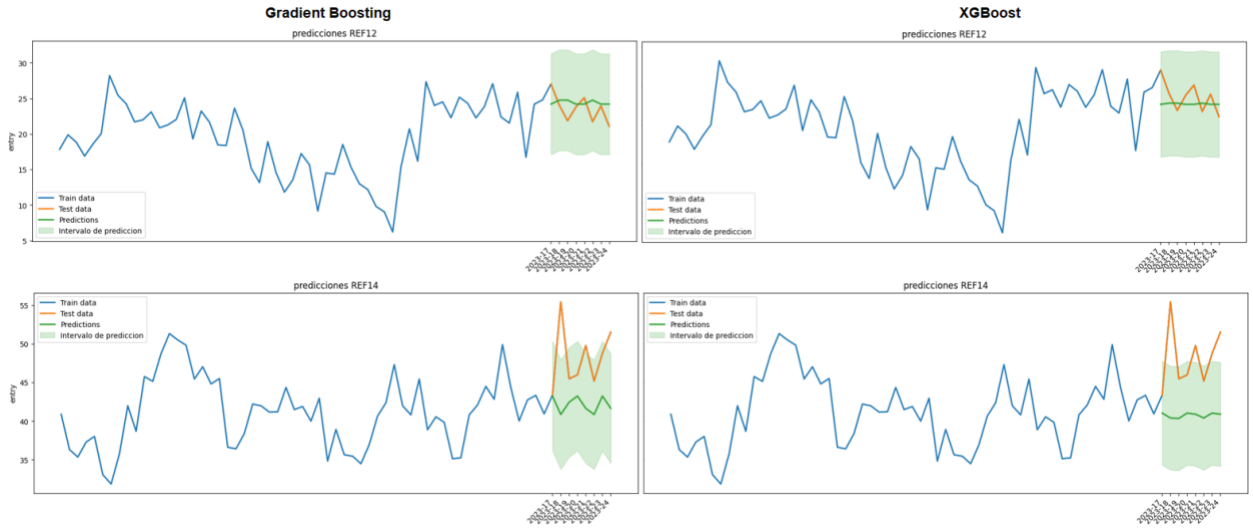

Figura 16: Intervalo de predicción REF12 y REF14

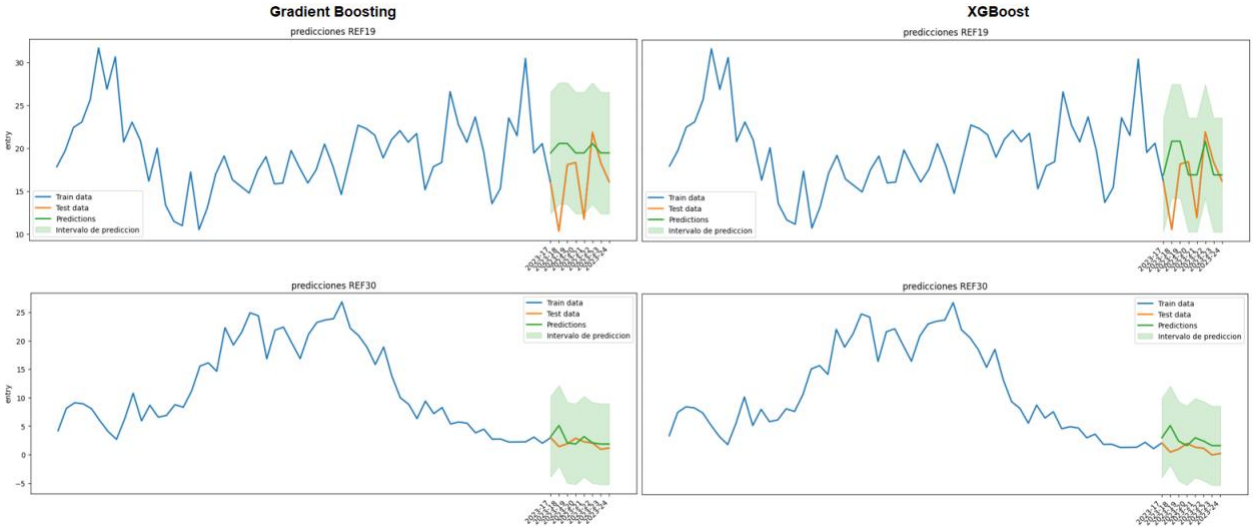

Figura 17: Intervalo de predicción REF19 y REF30

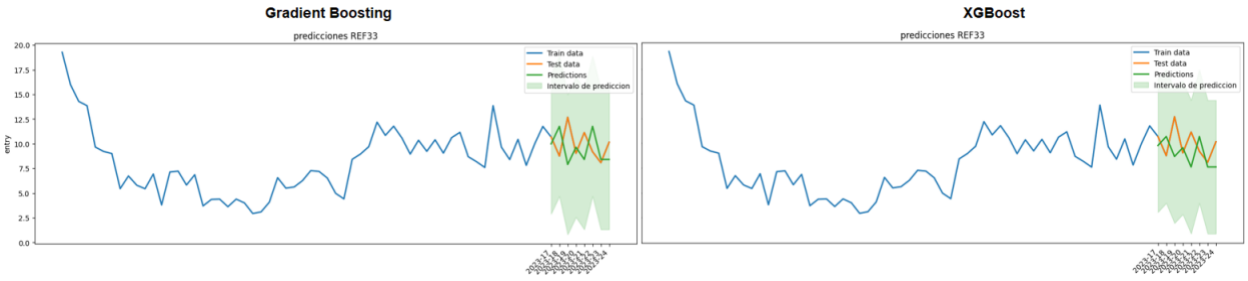

Figura 18: Intervalo de predicción REF33

Con la definición de intervalos de predicción a través de MapieRegressor se observa un rango de predicción que reduce la incertidumbre del modelo y que en la práctica

permitirá al equipo comercial tener un rango más amplio de acción en la toma de decisiones de inventarios o estrategias de penetración de mercado.

Por último, se busca interpretar el impacto de las variables explicativas sobre la variable explicada para estos dos modelos, este resultado se observa en la Figura 19:

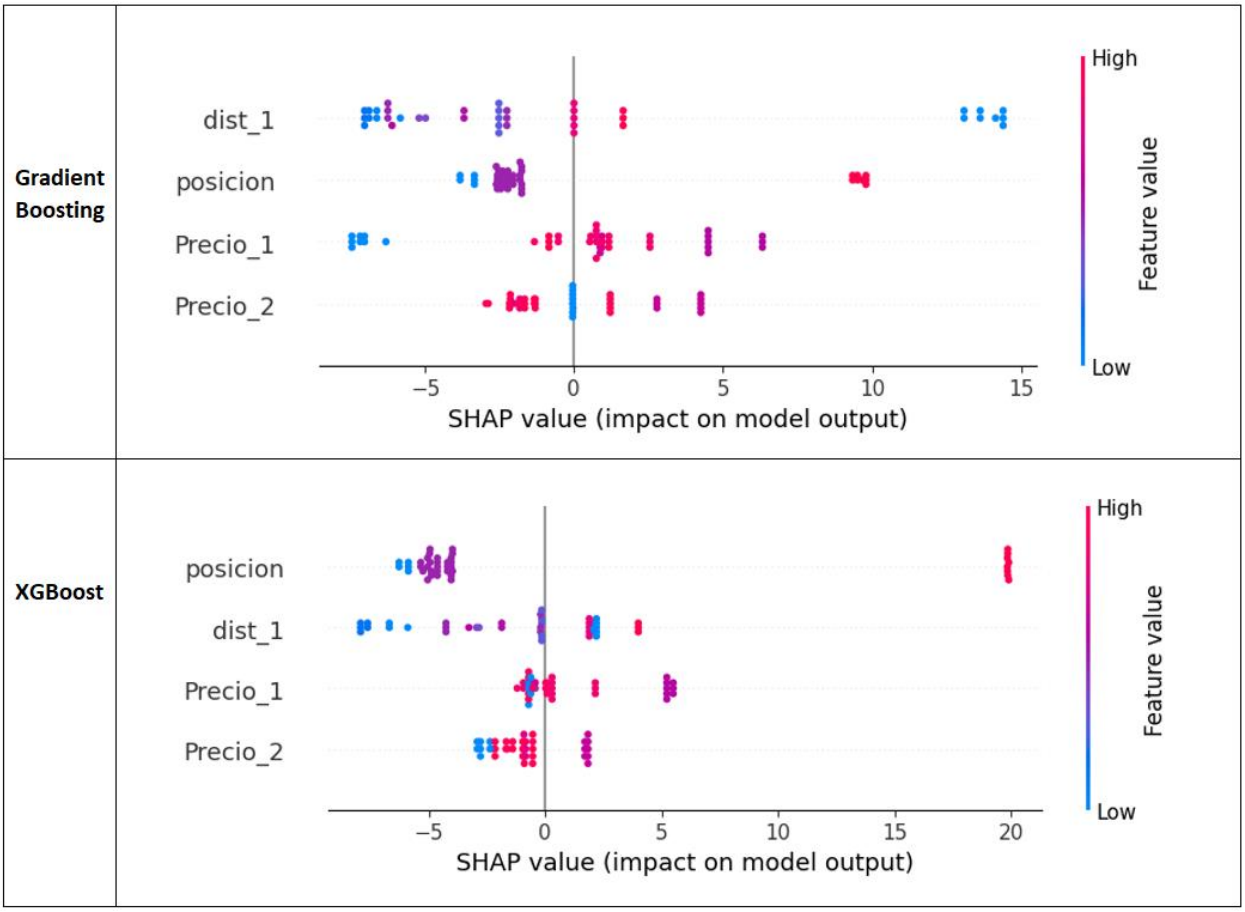

Figura 19: Impacto de las variables explicativas sobre la variable explicada para los modelos seleccionados para el set de pruebas

El gráfico SHAP para Gradient Boosting y XGBoost de la Figura 19 permite explicar la salida del modelo en términos de las variables explicativas con las que se entrenó, representando la contribución de cada una de estas. Para el modelo Gradient Boosting la característica más importante es la distancia, mientras que para XGBoost es la posición de la referencia frente a la línea VEL. Es importante observar que para ambos modelos las variables construidas según las definiciones del “Customer Value Map” son las que más contribuyen en el resultado final. Para ambos modelos en la Figura 19 un mayor valor de la variable posición aumenta la predicción de la variable Market Share.

Dados los resultados obtenidos, el modelo que minimiza el MAE para el segmento y que además tiene la mejor predicción a nivel de referencia es el Gradient Boosting. Para este modelo el MAE es de 2.94 puntos de Market Share.

6. Conclusiones

Las teorías del consumidor basadas en la metodología “Customer Value Map” pueden ser integradas a los modelos de predicción de ventas de motos nuevas. La mejor manera de integrar estas variables es a través de las definiciones planteadas alrededor de la línea de referencia VEL. Estas variables son distancia de cada referencia de moto a la línea VEL y la posición de la referencia en el “Customer Value Map”.

Con este estudio se identificó que existe una correlación negativa entre la distancia de la referencia de la moto a la línea VEL y el Market Share. Además, se logró construir varios modelos de Machine Learning basados en árboles que integran estas variables de teorías del consumidor con los precios de lista de las motos depurados de los bonos y promociones. Se encontró que el conocimiento del sector es muy importante para la detección de datos atípicos tales como: combinaciones de marca y segmento que no corresponden; periodos de tiempo con situaciones particulares en la industria; referencias que no deberían ser analizadas dentro del segmento. También, se concluye que los modelos de Machine Learning basados en árboles tienen un performance similar para este tipo de problemas de predicción del mercado de motos en Colombia (la diferencia en MAE es menor que 0.5 respecto al mejor modelo), siendo Gradient Boosting el que mejor se ajusta a las necesidades del negocio con un MAE de 2.94 puntos de Market Share y una serie de predicción que se ajusta a todas las referencias del segmento.

Con este estudio también se logró hallar un intervalo de confianza de la predicción que reduce la incertidumbre del modelo y permitirá a la empresa tomar decisiones basadas en un rango.

Vale la pena mencionar, que le modelo resultante de este estudio fue expuesto en la empresa que proporcionó los datos contando con la participación de los líderes comerciales y de marketing. Se decidió poner a competir los resultados del modelo con los pronósticos de series de tiempo previamente construidos por la empresa. Actualmente se están tomando decisiones con ambas metodologías buscando llegar a la mejor manera de pronosticar las ventas.

Finalmente, se puede concluir que haber desarrollado este modelo permitirá a las empresas del mercado de motos nuevas en Colombia tener una herramienta de predicción que se ajusta al dinamismo del mercado ya que el “Customer Value Map” se integra en las variables explicativas del modelo.

7. Entregable

Los resultados de este análisis se pueden consultar en el repositorio https://github.com/eludisa/Forecasting_ValueMap

8. Aspectos éticos

Los datos son de propiedad de la empresa a la cual se encuentra vinculado el estudiante, por disposición de esta no se revelará su nombre ni el nombre de las referencias de motos asociadas.

Los datos serán usados para la construcción de modelos que permitan alcanzar el objetivo planteado para el proyecto. Con el uso de estos datos, la empresa obtendrá un modelo que le permitirá tener información valiosa para sus procesos comerciales y productivos.

Para garantizar que se cumplan con las disposiciones de la empresa, los *dataset* compartidos con el profesor y con la universidad tendrán un proceso de anonimización previa.

9. Referencias

Acquila-Natale, E., & Iglesias-Pradas, S. (2021). A matter of value? Predicting channel preference and multichannel behaviors in retail. *Technological Forecasting and Social Change*, *162*. https://doi.org/10.1016/j.techfore.2020.120401

Bohanec, M., Kljajić Borštnar, M., & Robnik-Šikonja, M. (2017). Explaining machine learning models in sales predictions. *Expert Systems with Applications*, *71*, 416–428. https://doi.org/10.1016/j.eswa.2016.11.010

Buttner, D., & Rabe, M. (2021). Sales Forecasting in the Electrical Industry - An Illustrative Comparison of Time Series and Machine Learning Approaches. *2021 9th International Conference on Traffic and Logistic Engineering, ICTLE 2021*, 69–78. https://doi.org/10.1109/ICTLE53360.2021.9525747

Dairu, X., & Shilong, Z. (2021). Machine Learning Model for Sales Forecasting by Using XGBoost. *2021 IEEE International Conference on Consumer Electronics and Computer Engineering, ICCECE 2021*, 480–483. https://doi.org/10.1109/ICCECE51280.2021.9342304

Deng, T., Zhao, Y., Wang, S., & Yu, H. (2021). Sales Forecasting Based on LightGBM. *2021 IEEE International Conference on Consumer Electronics and Computer Engineering, ICCECE 2021*, 383–386. https://doi.org/10.1109/ICCECE51280.2021.9342445

Di Pillo, G., Latorre, V., Lucidi, S., & Procacci, E. (2016). An application of support vector machines to sales forecasting under promotions. *4OR*, *14*(3), 309–325. https://doi.org/10.1007/s10288-016-0316-0

Gautam, N., Nayak, S., & Shebalov, S. (2021). Machine learning approach to market behavior estimation with applications in revenue management. *Journal of Revenue and Pricing Management*, *20*(3), 344–350. https://doi.org/10.1057/s41272-021-00317-y

Hern Kong, Y., Yin Lim, K., & Yoke Chin, W. (2021). *Time Series Forecasting Using a Hybrid Prophet and Long Short-Term Memory Model* (A. Mohamed, B. W. Yap, J. M. Zain, & M. W. Berry, Eds.; Vol. 1489). Springer Singapore. https://doi.org/10.1007/978-981-16-7334-4

IBM. (2020). *Conceptos básicos de ayuda de CRISP-DM - Documentación de IBM*. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

Jiang, H., Ruan, J., & Sun, J. (2021). Application of Machine Learning Model and Hybrid Model in Retail Sales Forecast. *2021 IEEE 6th International Conference on Big Data Analytics, ICBDA 2021*, 69–75. https://doi.org/10.1109/ICBDA51983.2021.9403224

Kiely, T. J., & Bastian, N. D. (2019). *The Spatially-Conscious Machine Learning Model*. http://arxiv.org/abs/1902.00562

Kolková, A., & Ključnikov, A. (2022). DEMAND FORECASTING: AI-BASED, STATISTICAL AND HYBRID MODELS VS PRACTICE-BASED MODELS-THE CASE OF SMES AND LARGE ENTERPRISES. *RECENT ISSUES IN ECONOMIC DEVELOPMENT*. https://doi.org/10.14254/2071

Kolková, A., & Navrátil, M. (2021). Demand forecasting in python: Deep learning model based on lstm architecture versus statistical models. *Acta Polytechnica Hungarica*, *18*(8), 123–141. https://doi.org/10.12700/APH.18.8.2021.8.7

Leszinski, R., & Marn, M. (2016). Setting value, not price. *McKinsey Quarterly*.

Marn, M. V, Roegner, E. V, & Zawada, C. C. (2004). *The Price Advantage*. www.WileyFinance.com.

Menculini, L., Marini, A., Proietti, M., Garinei, A., Bozza, A., Moretti, C., & Marconi, M. (2021). Comparing Prophet and Deep Learning to ARIMA in Forecasting Wholesale Food Prices. *Forecasting*, *3*(3), 644–662. https://doi.org/10.3390/forecast3030040

Mohd, R. S., Suhardi, W. M., Anita, A. H., Maznah, W. O., & Etty, H. H. (2013). The relationship between product quality and purchase intention: The case of Malaysias national motorcycle/scooter manufacturer. *African Journal of Business Management*, *5*(20), 8163–8176. https://doi.org/10.5897/ajbm11.267

Navratil, M., & Kolkova, A. (2019). Decomposition and forecasting time series in business economy using prophet forecasting model. *Central European Business Review*, *8*(4), 26–39. https://doi.org/10.18267/j.cebr.221

Panarese, A., Settanni, G., Vitti, V., & Galiano, A. (2022). Developing and Preliminary Testing of a Machine Learning-Based Platform for Sales Forecasting Using a Gradient Boosting Approach. *Applied Sciences (Switzerland)*, *12*(21). https://doi.org/10.3390/app122111054

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. *Data*, *4*(1). https://doi.org/10.3390/data4010015

Rasim, Junaeti, E., & Wirantika, R. (2018). Implementation of Automatic Clustering Algorithm and Fuzzy Time Series in Motorcycle Sales Forecasting. *IOP Conference Series: Materials Science and Engineering*, *288*(1). https://doi.org/10.1088/1757-899X/288/1/012126

Sasmita, Y., & Darmawan, G. (2017). Accuracy evaluation of Fourier series analysis and singular spectrum analysis for predicting the volume of motorcycle sales in Indonesia. *AIP Conference Proceedings*, *1868*. https://doi.org/10.1063/1.4995125

Sharif Azadeh, S., Marcotte, P., & Savard, G. (2015). A non-parametric approach to demand forecasting in revenue management. *Computers and Operations Research*, *63*, 23–31. https://doi.org/10.1016/j.cor.2015.03.015