UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

"DESARROLLO DE UN MODELO DE PREDICCIÓN DE TRÁFICO DE CLIENTES PARA OPTIMIZAR LA DOTACIÓN DE PERSONAL EN TIENDAS FÍSICAS DEL *RETAIL*"

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

MATÍAS ARTURO CHRISTIANSEN AGAR

PROFESOR GUÍA:
MARCELO OLIVARES ACUÑA

MIEMBROS DE LA COMISIÓN:
VICTOR BUCAREY LÓPEZ
DANIEL YUNG MEYOHAS

SANTIAGO, CHILE
2020

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial
POR: Matías Arturo Christiansen Agar
FECHA: 18/06/20
PROFESOR GUÍA: Marcelo Olivares Acuña

DESARROLLO DE UN MODELO DE PREDICCIÓN DE TRÁFICO DE CLIENTES PARA OPTIMIZAR LA DOTACIÓN DE PERSONAL EN TIENDAS FÍSICAS DEL *RETAIL*

El presente trabajo corresponde al desarrollo de un modelo de predicción robusto de tráfico de clientes en tiendas de *retail* físicas en Chile, con el fin mejorar la rentabilidad de éstas, a través de una dotación de personal óptima que se ajuste al tráfico y minimice la sobre y sub dotación. Esta motivación surge producto de que en los últimos años la industria del *retail* físico ha bajado sus rentabilidades y el aumento del *big data* permite implementar soluciones tecnológicas con la finalidad de entender el comportamiento del cliente.

Para la obtención de una mejor predicción, se realizan cuatro actividades claves. En primer lugar, se identifican variables exógenas, que puedan explicar la fluctuación en el tráfico de clientes. Las principales variables identificadas son los días feriados, separados por tipo (renunciable o irrenunciable); y la precipitación, que afecta negativa o positivamente dependiendo de si la tienda está ubicada en la calle o en un *mall* respectivamente. En segundo lugar, se escoge una herramienta de predicción robusta, como Prophet, que pueda ser usado en distintas tiendas de *retail*, independiente de la especialidad. En tercer lugar, se realiza un control estadístico de procesos para identificar cuando los sensores de medición fallan. Finalmente, se lleva a cabo una imputación de datos de alta precisión (MAPE = 15.24%), con el fin de minimizar el ruido de la serie temporal.

Gracias a esta investigación, se concluye que Prophet es una herramienta poderosa y robusta de fácil uso. El modelo desarrollado incluye estacionalidad semanal, mensual y anual, como también regresores que captan el efecto de los días feriados y la lluvia. Se obtiene que el MAPE promedio de este modelo es de 25.70%, en comparación al 55.21% del modelo desarrollado por Lam et al. (1998). Dicha diferencia en la precisión se puede traducir en un incremento de la rentabilidad del 1.45%, correspondiente a la optimización de la dotación de trabajadores en las tiendas.

Dedicatoria

A mis padres, Guillermo y Ximena.

Agradecimientos

Quisiera agradecer a Intelligenxia por darme la oportunidad y flexibilidad de desarrollar la presente memoria. En específico, a Nicolás, Raúl y Denise, los cuales dispusieron siempre de su conocimiento y generaron un agradable ambiente de trabajo.

También quiero agradecer a mi profesor guía Marcelo Olivares, el cual hizo todo esto posible y me otorgó la flexibilidad para escribir la memoria desde el extranjero, y a los miembros de la comisión Víctor Bucarey y Daniel Yung. Muchas gracias.

A mis amigos que me acompañaron en esta larga travesía en Beauchef. Diego (en especial por todos los trabajos que hicimos), Gastón, Sebastián, Agustín, Benjamín, Vicente y Cristóbal. Gracias por todo. Jamás olvidaré las maratones de CAA en la casa del coloro, jamás olvidaré cuando nos conocimos todos bien en Métodos experimentales con Dulic y obviamente no olvidaré el tremendo viaje a Cancún gracias a nuestro patrocinador, Aeroméxico. En fin, jamás los olvidaré, mechones.

Asimismo, quiero agradecer a mi familia por siempre ser un apoyo, en particular a mis padres. Viejos, espero que estén tan contentos como yo, y sobre todo ahora que, ¡Oficialmente somos colegas e hijos(as) de bello! También agradecer a mis hermanos, Karen y Alex por siempre alegrarme los días y compartir hermosos momentos juntos. A la Mara, por su infinita felicidad y obviamente a ti, Greg, por siempre estar ahí en las buenas y en las malas. Gracias totales.

Imposible no agradecerte a ti, Giulia. Gracias por ser una inspiración constante en mi vida. Te quiero agradecer por todos los momentos de alegría que hemos compartido estos 2 años. A veces me pongo a pensar ¿Qué hubiera pasado si no te hubiera contado ese chiste en la clase? Las cosas definitivamente ocurren por algo y estoy muy feliz de poder estar a tu lado. Igualmente, muchas gracias por confiar en mí y apoyarme durante todo este eterno proceso. *Alò. Grazie!*

Tabla de contenido

Índice de tablas

Índice de figuras

1. Introducción

1.1. Identificación del problema u oportunidad

Luego de un extenso crecimiento de dos dígitos entre 2011 y 2014, también distinguido como el super ciclo del consumo, la industria del *retail* físico en Chile transitó hacia una etapa de desaceleración, acompañado también de la baja actividad económica en general del país, el cual ha afectado la creación de empleos, los ingresos de los hogares y, en consecuencia, las expectativas de los consumidores.

Sin embargo, aún con el problema de una industria debilitada, el *retail* físico ha presenciado grandes transformaciones. Principalmente, la inversión se ha focalizado en elevar la eficiencia, márgenes y productividad, en vez de ampliar la cantidad de tiendas o superficie.

En consecuencia, las tiendas físicas deben reinventarse, construyendo ambientes para crear negocio más allá de la simple venta del producto. “En la actualidad, las tiendas y en general cualquier negocio con espacios físicos de atención al público, están modificando la forma de hacer gestión” (Intelligenxia, 2019). Actualmente, las tiendas tradicionales están migrando a un modelo de negocios basado en los datos (*data driven*). Este modelo permite principalmente predecir el tráfico futuro, lo cual es esencial para generar algoritmos que recomiende nuevos productos a sus clientes y la personalización de promociones u ofertas, para así poder competir con este nuevo canal de compra.

Recientes innovaciones tecnológicas permiten pronosticar de forma más precisa el tráfico futuro, a través de la demanda de clientes a una tienda. “Con la utilización de sensores 3D, la información sobre el tráfico de clientes tanto en el exterior e interior de la tienda fluye al instante al contar con Wi-Fi” (Intelligenxia, 2019).

Dado lo anterior, el propósito del presente trabajo de título es brindar algunas soluciones a las empresas de *retail*. En particular, se busca mejorar las predicciones de una tienda (perteneciente a una cadena de tiendas) a través de un modelo robusto, como también identificando el ruido asociado a la falla de los sensores. Para el desarrollo del trabajo, se utilizarán datos históricos de 54 tiendas a lo largo de Chile, pertenecientes a tres tipos de tiendas: repuestos automovilísticos (18), vestimenta infantil (13) y vestimenta deportiva (23).

1.2. Causales del problema identificado

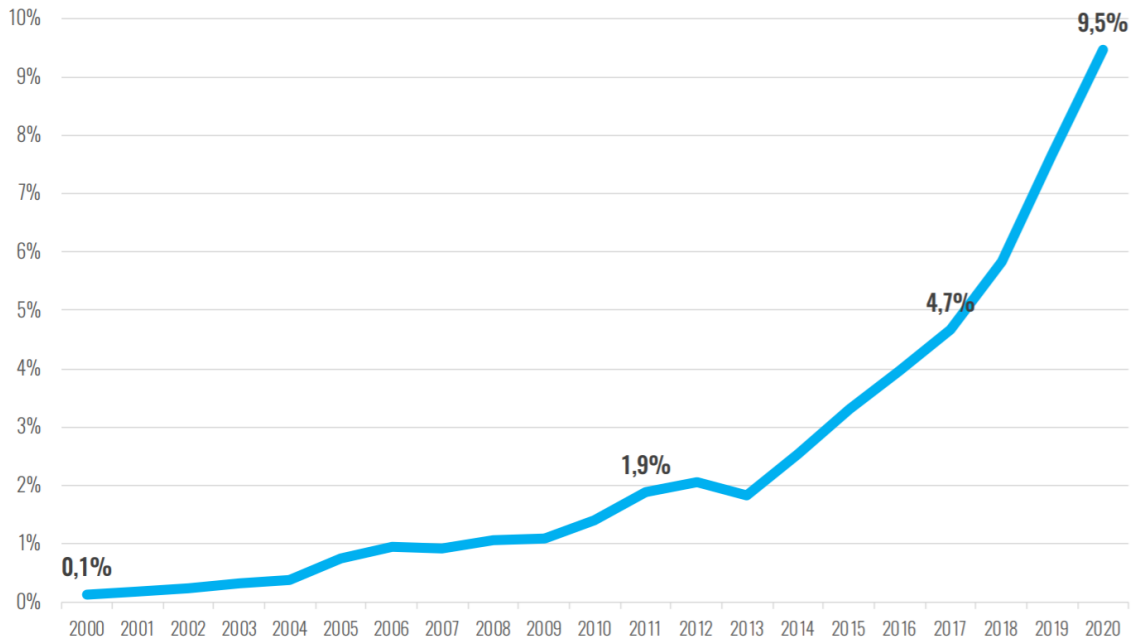
Con motivo de identificar las causas del problema encontrado, la Cámara de Comercio de Santiago (CCS) menciona los diferentes factores influyentes en la desaceleración del *retail* físico, los cuales se mencionan a continuación.

- Crecimiento del *retail* online
- Aumento de incidencia de la tecnología asociado a la fidelización del cliente
- Aumento de la competencia internacional

En primer lugar, debido al mundo dinámico y globalizado de la actualidad, la industria del *retail* ha sufrido un enorme cambio en las últimas dos décadas. El *retail online* (*e-commerce*) ha aumentado exponencialmente en Chile, siguiendo la tendencia mundial. Según la Cámara de Comercio de Santiago (CCS), se espera que la penetración del *retail online* alcance el 9.5% de la industria (***Figura 1***).

Figura 1

Penetración del retail online sobre el total en Chile.



Fuente: Centro Economía Digital CCS: *"Tendencias del comercio electrónico en Chile"*.

Bajo este contexto, el *retail* físico (*brick-and-mortar*) presenta una gran amenaza por parte del *e-commerce*, en términos de cuota de mercado. Sin embargo, las tiendas físicas no siempre son competidores, sino que totalmente lo contrario, es decir, son complementarios. Lo anterior se debe a que la mayoría de las tiendas de *e-commerce* actualmente poseen tiendas físicas. Además, Según un estudio realizado por la CCS en 2019, la primera barrera identificada de los chilenos es probar los productos antes de comprarlos (***Figura 1 del Anexo I***).

En segundo lugar, con el objetivo de fidelizar a los clientes – y en conjunto con la gran capacidad de procesar datos – las empresas están invirtiendo en crear experiencias. Sin embargo, existe un gran costo asociado en la obtención de datos (por ejemplo; la compra, instalación y mantención de sensores) y en el análisis de éstos, a través de la contratación de *data scientists*. Dichos costos y el enfoque en la experiencia han reducido la tasa de crecimiento experimentada por la industria en el super ciclo.

Finalmente, varias empresas de retail han ingresado al mercado chileno en busca de crecimiento en mercados emergentes. Dentro de dichas empresas se encuentra H&M, que inicio operaciones en 2013 y al momento de la escritura, posee 18 tiendas a lo largo de Chile. Asimismo, la globalización y el *e-commerce* han potenciado la entrada de mega actores virtuales como lo son Amazon, eBay y AliBaba, entre otros. Dichos *retailers* presentan una ventaja competitiva en el pago de impuestos de consumo, debido a su naturaleza virtual.

1.3. Valor agregado mediante la investigación

La presente investigación propone un modelo de predicción de tráfico robusto que sea preciso para todo tipo de tiendas de *retail*. Lo interesante de este modelo es que no es necesario tener un conocimiento profundo en pronóstico de series de tiempo para obtener predicciones robustas.

Además, predicción de tráfico conlleva a diversos ahorros en caso de que ésta sea precisa. En primer lugar, tener un mejor conocimiento de la demanda futura puede optimizar la cantidad de inventario, reduciendo dichos costos. En segundo lugar, al poder estimar la demanda, se posee más información para que los gerentes de tiendas puedan tomar decisiones operativas y tácticas. Finalmente, se puede optimizar la dotación de personal en tienda, evitando la sobre o sub dotación. Es imprescindible mencionar que el gasto asociado a la fuerza de trabajo en la industria de *retail* se estima del orden del 12% de las ventas.

La sobre dotación implica un gasto innecesario al tener personal trabajando a una baja capacidad debido a la baja demanda. Usualmente se identifica una sobre dotación de vendedores en los horarios de bajo tráfico (por ejemplo, en la mañana). Por el otro lado, la sub dotación implica ventas potenciales que son perdidas al no tener un trabajador que oriente al cliente y completar la venta.

Finalmente, se presenta un algoritmo que detecta cuando los sensores de tráfico están fallando, con la finalidad de arreglarlos lo antes posible e ignorar dichas mediciones al ser datos errados.

1.4. Objetivos

1.4.1. Objetivo general

El objetivo general del trabajo de título consiste en desarrollar un modelo de predicción de tráfico de clientes en una tienda física de *retail*, con el fin de realizar una dotación de personal más precisa, minimizando la sub y sobre dotación.

1.4.2. Objetivos específicos

Para poder cumplir objetivo general, se proponen los siguientes objetivos específicos, los cuales indican lo que se pretende realizar en cada etapa de investigación, y al ser alcanzado en conjunto, permite garantizar el cumplimiento del objetivo general:

1) Realizar un análisis exploratorio y determinar las variables exógenas a incluir en el modelo.

2) Establecer un modelo de imputación para los datos faltantes (*missing values*)

3) Estimar los distintos modelos predictivos e identificar el más robusto en base a la métrica de precisión seleccionada

4) Calcular el ahorro promedio por tienda asociado a la nueva dotación de personal

5) Establecer un control estadístico de procesos de los sensores y obtener un reporte que indica cuándo fallan.

2. Marco teórico

En esta sección, se detallan diferentes conceptos esenciales para la total comprensión de los distintos modelos de pronóstico, como también las métricas de precisión para clasificarlos. Finalmente se expone las distintas tecnologías utilizadas para poder recopilar los datos de tráfico.

A grandes rasgos, las áreas de la ciencia con mayor relevancia en el desarrollo del presente trabajo corresponden a Gestión de Operaciones y *Machine Learning*, ya que, ambas conforman la base para comprender el contexto y los factores influyentes en el tema a abordar, correspondiente al pronóstico de series de tiempo y dotación de personal. Junto a ellas, se destacan también como significativas las áreas de Estadística y de Optimización. La primera, debido a la importancia del procesamiento de datos al momento de realizar cualquier análisis y, la segunda, relacionada al uso de simulaciones para obtener las utilidades asociadas a una dotación de personal.

2.1. Definición y clasificación de una serie temporal

Las series temporales es una secuencia de datos numéricos en orden sucesivo que son utilizadas para entender la relación causal entre variables que cambian a lo largo del tiempo. En otras palabras, una serie de tiempo puede ser cualquier variable que cambie con el tiempo. Las series de tiempo se pueden desglosar en distintos componentes:

1. Tendencia (T_t): Corresponde a la tendencia general de los datos (que puede aumentar o disminuir) durante un largo período de tiempo. Una tendencia es suave, general y a largo plazo. No siempre es necesario que el aumento o la disminución estén en la misma dirección durante todo el período de tiempo dado. Es posible que la serie de tiempo puede aumentar, disminuir o ser estable en distintas secciones del tiempo. Sin embargo, la tendencia general debe ser al alza, a la baja o estable. La tendencia puede ser lineal o no lineal.

2. Estacionalidad (E_t): Corresponde a un movimiento periódico debido a la influencia de factores exógenos que se repiten a lo largo de la serie temporal. Usualmente corresponde a días, semanas, meses y años.

3. Ruido (ε_t): Es un factor que causa la variación aleatoria e irregular en la variable en estudio. Este componente es incontrolable e impredecible.

Con los componentes definidos, existen tres tipos de series temporales:

1. Aditivas: Se componen sumando todas las componentes anteriores. En este caso, la tendencia es lineal.

$$x_t = T_t + E_t + \varepsilon_t$$

2. Multiplicativas: Se componen multiplicando todas las componentes anteriores. En este caso, la tendencia no es lineal.

$$x_t = T_t \times E_t \times \varepsilon_t$$

3. Mixtas: Corresponde a una mezcla sumas y multiplicaciones de las componentes. Existen varias combinaciones posibles como las siguientes.

$$x_t = T_t + E_t \times \varepsilon_t$$
$$x_t = T_t \times E_t + \varepsilon_t$$

2.2. Modelos de predicción

Al momento de realizar predicciones, no existe un único que sea superior. De hecho, existe una gran variedad de modelos para efectuar estimaciones, donde cada uno posee sus fortalezas y debilidades. El rol del investigador es poder determinar correctamente el modelo a utilizar con sus respectivos parámetros.

Los modelos de series de tiempo tienen un objetivo netamente predictivo y los pronósticos obtenidos se generan basados al comportamiento pasado de la variable a predecir. Es posible distinguir dos tipos de modelos de series de tiempo: Modelos deterministas y modelos estocásticos. Los modelos deterministas son caracterizados por su simplicidad, debido a que no consideran la aleatoriedad subyacente en la serie, lo cual implica una baja precisión en las predicciones. En cambio, los modelos estocásticos consideran la aleatoriedad, por lo que se generan modelos aproximados que son efectivos para la generación de pronósticos.

Además, los modelos se basan en el supuesto que las series de tiempo son estacionarias. Las series estacionarias son aquellas que la media y varianza se mantienen fijas durante todo el tiempo, mientras que la covarianza es función del rezago. Por el otro lado, las series no estacionarias consideran la media, varianza y covarianza como variables por lo que cambian con el tiempo. Lo anterior implica que las series no estacionarias sean más complejas de modelar. A continuación, se detallan los principales modelos utilizados para pronosticar una serie de tiempo:

2.2.1. ARIMA

El modelo autorregresivo integrado de promedio móvil (ARIMA) utiliza variaciones y regresiones de datos estadísticos con el fin de obtener patrones intrínsecos de la serie para luego predecir el futuro. Se suele expresar de manera general como $ARIMA(p, d, q)$ donde los parámetros p, d y q son los distintos componentes del modelo:

- Autorregresiva (AR): Asume que el valor de la serie de tiempo en t corresponde a una combinación lineal de valores perteneciente a los p periodos anteriores.

- Integrada (I): Aplica diferenciación en los datos con el objetivo de obtener una serie estacionaria. El parámetro d corresponde a la cantidad de diferenciaciones realizadas.

  Para diferenciar la serie de tiempo, se calcula la diferencia entre dos observaciones consecutivas, eliminando la tendencia y estacionalidad, obteniendo una serie estacionaria:

  $$y'_t = y_t - y_{t-1}$$

  Hay casos que es necesario realizar más de una diferenciación para obtener una serie estacionaria.

- Media Móvil (MA): Asume que el error de regresión observado corresponde a una combinación lineal de q errores aleatorios previos.

El modelo ARIMA se desarrolló principalmente en a fines de la década de 1960 y posteriormente fue sistematizado por Box y Jenkins (1976) con los siguientes pasos fundamentales:

1. Por medio de transformaciones y/o diferencias se estabiliza la varianza y se elimina la tendencia y la estacionalidad de la serie, obteniéndose así una serie estacionaria.

2. Para la serie estacionaria obtenida se identifica y se estima un modelo que explica la estructura de correlación de la serie con el tiempo.

3. Al modelo hallado en el punto 2, se le aplican transformaciones inversas que permiten establecer la variabilidad, la tendencia y la estacionalidad de la serie original.

4. El modelo estimado se valida a través de la correlación de sus residuales. Si estos llegan a presentar correlación es necesario volver a estimar nuevamente los parámetros, es decir, se regresa al punto 2. El anterior procedimiento iterativo se repite hasta que finalmente no exista correlación significativa entre los residuales.

El modelo ARIMA se caracteriza principalmente por ser un modelo técnicamente sofisticado para la predicción de una variable. Analiza errores recientes de pronósticos para luego seleccionar el ajuste más apropiado para el futuro y utiliza la observación más reciente como valor inicial. Además, es un modelo más apropiado para predicciones a largo plazo.

La ***Tabla 1*** destaca las ventajas y desventajas del modelo ARIMA (Thomas, 1983).

Tabla 1

Ventajas y desventajas del modelo ARIMA.

Ventajas	Desventajas
El modelo se ajusta a una serie en particular.	Requiere de una serie temporal estacionaria.
Considera efectivamente la correlación lineal entre observaciones	La precisión depende de la experiencia del pronosticador.

Fuente: Creación propia.

2.2.2. Árboles Aleatorios

El modelo ARIMA es bastante flexible, pero está limitado por el supuesto de dependencias intertemporales lineales. Por consiguiente, no son capaces de capturar efectos no lineales, que a menudo están presentes en el mundo real (Zhang et al., 1998, 2001; Gooijer y Hyndman. 2006).

Una variedad de modelos han surgido buscando capturar los efectos no lineales. Unos ejemplos son el modelo bilineal (Granger y Anderson, 1978), modelo STAR (Chan y Tong, 1986), modelo ARCH (Engle, 1982) y el modelo GARCH (Bollerslev, 1986; Taylor, 1987), entre otros. Lamentablemente, la mayoría de estos modelos funcionan bien para problemas específicos, pero no logran generalizar para otras series temporales no lineales.

La insuficiencia predictiva y el crecimiento de los procesadores computacionales han impulsado el crecimiento de los modelos de *machine learning*. La gran ventaja de estos modelos, sobre los modelos clásicos no lineales, es que no requieren supuestos sobre la estructura preliminar de los datos (Ej. estacionalidad).

Árboles aleatorios (Breiman, 2001), también conocido como *random forests* en inglés, es una técnica que promedia las predicciones de una gran cantidad de árboles de decisión no correlacionados entre ellos. Usualmente se obtiene un buen rendimiento con mejores propiedades de generalización que los árboles de decisión, son relativamente robustos a los *outliers* y prácticamente no requieren ajuste de parámetros (Raschka, 2015).

La ***Tabla 2*** destaca las ventajas y desventajas del modelo *random* forest (Kumar, 2019).

Tabla 2

Ventajas y desventajas del modelo de árboles aleatorios.

Ventajas	Desventajas
Uno de los modelos más robustos que existen actualmente.	Dificultad de interpretar los resultados.
Gran capacidad de procesamiento (cantidad de datos y variables).	En algunos casos, el modelo sobre ajusta provocando pronósticos ruidosos.
Buen manejo de *outliers*. Funciona bien aún con varios *missing values*.	Gran costo de procesamiento computacional y tiempo.

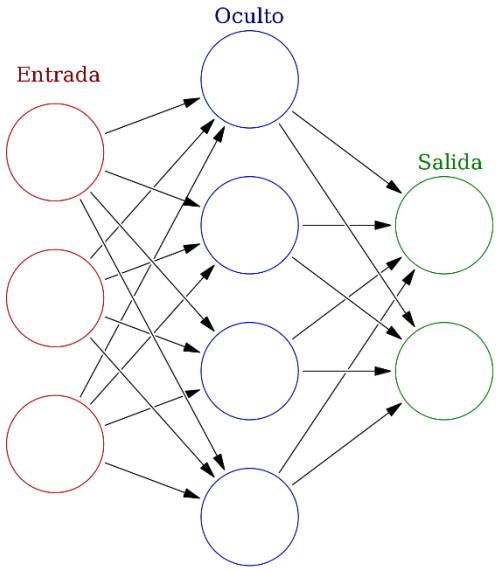
Fuente: Creación propia.

2.2.3. Redes neuronales

Una red neuronal se puede describir como un modelo de regresión no lineal cuya estructura se inspira en el funcionamiento del sistema nervioso. En términos generales, una red neuronal consiste en un gran número de unidades simples de proceso, denominadas neuronas, que actúan en paralelo. Además, estas neuronas están agregadas en capas y están conectadas mediante vínculos ponderados (Ríos, 2008). La ***Figura 2*** representa gráficamente el concepto de una red neuronal.

Figura 2

Ejemplo de una red neuronal artificial.



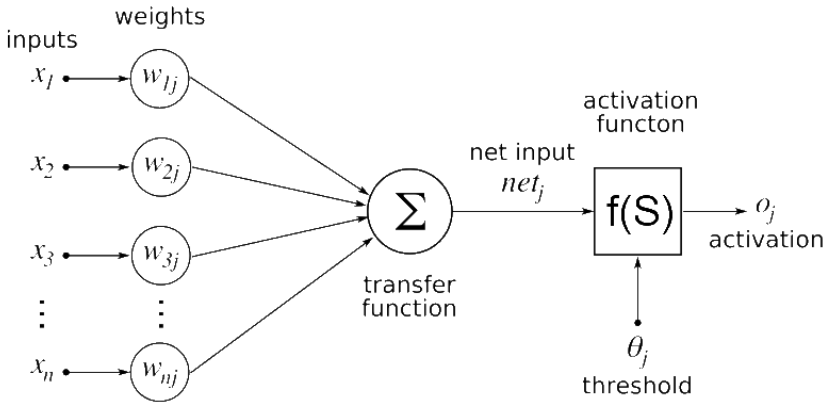
Fuente: Recuperado de *"Red neuronal artificial"*, Wikipedia.

Al igual que en el cerebro, cada neurona recibe distintos *inputs* desde otras neuronas y genera un resultado que depende solo de la información localmente disponible, ya sea almacenada internamente o plasmada en los ponderadores de las conexiones. Finalmente, el *output* generado por la neurona servirá de *input* para otras, formándose una red de neuronas.

La ***Figura 3*** muestra el funcionamiento de una neurona.

Figura 3

Funcionamiento de una neurona en una red neuronal.



Fuente: Recuperado de *"Red neuronal artificial"*, Wikipedia.

La llamada función de activación es una función que emula el umbral presente en el sistema nervioso. Esto significa si la respuesta de una neurona no es lo suficientemente grande, entonces esta no afecta en las siguientes neuronas. Las funciones más usadas son la función de escalón, signo, sigmoidal, gaussiana y lineal.

Al igual que el modelo de *random forest*, los algoritmos de redes neuronales pueden aprender por si mismos a medida que se le proporciona nueva información, el cual adapta las ponderaciones de las conexiones entre nodos.

Es importante destacar que, entre mayor cantidad de capas, la red neuronal puede incorporar más información. En particular, un modelo sin capas ocultas es equivalente a un modelo de regresión lineal. Lo anterior se debe a que los valores de salida corresponden a una combinación lineal de los valores de entrada.

La ***Tabla 3*** resume las ventajas y desventajas de las redes neuronales (Mijwil, 2018).

Tabla 3

Ventajas y desventajas del modelo de redes neuronales.

Ventajas	Desventajas
Tolerancia a fallos debido a que almacena la información de forma redundante.	Aumento exponencial de la complejidad al incrementar la flexibilidad / cantidad de variables.
Flexibilidad ya que puede manejar cambios leves en la información de entrada sin modificar el resultado final.	No permite interpretar lo que se ha aprendido. La red neuronal solamente proporciona un valor de salida.
Se obtienen respuestas en tiempo real debido a la estructura paralela.	Falta de reglas definitorias para construir una red para un problema determinado.

Fuente: Creación propia.

2.2.4. Prophet

En febrero 2017, Facebook liberó una herramienta (de uso público) llamada Prophet. Dicha herramienta es una librería que permite construir modelos de ajuste y pronóstico de series. Para ello no utiliza los métodos más tradicionales mencionados anteriormente como ARIMA, sino que lo que Facebook denomina *curve fitting*.

Prophet permite manejar sets de datos donde existan observaciones nulas, faltantes, *outliers*, cambios de tendencia importantes (por ejemplo, las semanas previas a Navidad) y tendencias no lineales. Es decir, resulta considerablemente versátil para utilizarlo en series de tiempo sobre la cual se quieran realizar predicciones. (López, 2018). Dicha herramienta es un modelo regresivo aditivo con las tres componentes de una serie temporal:

1. Tendencia: Detecta automáticamente cambios de tendencia seleccionando los distintos quiebres de tendencia dentro del set de datos. Así arma una función (la cual está definida por partes) de tendencia linear o de crecimiento logístico (que alcanza nivel de saturación).

$$g(t) = kt \qquad g(t) = \frac{C}{1 + e^{k(m-t)}}$$

   donde C es el nivel de saturación, k es el *ratio* de crecimiento y m es un parámetro *offset*.

2. Estacionalidad: Se modela utilizando series de Fourier con el fin de obtener un modelo más flexible. Los efectos estacionales se determinan a través de la siguiente función:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right)$$

3. Ruido: El término de error ε_t representa cualquier cambio idiosincrático que no puede ser acomodado por el modelo. Se asume que dicho termino distribuye normalmente.

Además de los tres componentes mencionados, Prophet permite un cuarto para días feriados o eventos, cuya finalidad es señalar hechos concretos que pueden alterar los valores de la serie.

En consecuencia, Prophet puede ser descrito por la siguiente suma aditiva de los componentes.

$$x_t = T_t + E_t + H_t + \varepsilon_t$$

Prophet actúa como una caja negra la cual se le proporciona una cantidad de *inputs* (serie de tiempo, variables independientes, estacionalidades, feriados o fechas especiales entre otros), el cual procesa la información y se obtiene el pronóstico. Además, indica los *outliers*, cambios de tendencia, impacto marginal de los feriados y recomendaciones para mejorar el modelo, entre otros.

La ***Tabla 4*** muestra las principales ventajas y desventajas del modelo Prophet (Taylor y Letham, 2017).

Tabla 4

Ventajas y desventajas del modelo de Prophet.

Ventajas	Desventajas
Fácil uso del modelo para principiantes. Analistas con experiencia fácilmente pueden extender el modelo para incluir o excluir diversos componentes.	Menor capacidad de procesamiento que árboles aleatorios.
Facilidad en la interpretación de los componentes	
Rapidez en el ajuste del modelo, lo que permite al analista explorar variaciones del modelo.	

Fuente: Creación propia.

2.3. Métricas de precisión

Al momento de generar predicciones de tráfico, es imprescindible medir la precisión de estas con el fin de decidir cuál es el mejor modelo predictivo. La medición de la predicción de pronóstico es una tarea compleja debido a que no existe un indicador único. A continuación, se establecen las métricas más utilizadas actualmente por la academia:

1. RMSE: La raíz cuadrática media del error (RMSE en inglés) mide la dispersión de una variable simulada en el transcurso del tiempo. El RMSE queda definido como:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{Y}_t - Y_t\right)^2}$$

   Esta métrica penaliza bastante los errores grandes ya que estos están elevados al cuadrado. Dado lo anterior, es útil utilizar dicha métrica cuando el costo de cometer un error es aproximadamente proporcional al cuadrado de dicho error.

2. MAPE: El promedio del error porcentual absoluto (MAPE) mide el valor medio del error absoluto en términos porcentuales al valor real de la variable. El MAPE queda definido como:

$$MAPE = \frac{1}{T}\left(\sum_{t=1}^{T}\frac{\left|\widehat{Y}_t - Y_t\right|}{Y_t}\right) \times 100$$

   Al expresar el error porcentual, el MAPE es un indicador utilizado frecuentemente debido a su facilidad en la interpretación de resultados. Asimismo, es útil aún cuando se desconoce el valor real, ya que se expresa de manera relativa a dicho valor.

   Además, cabe destacar que no siempre el modelo que genere pronósticos con un menor tendrá un menor RMSE y viceversa. Lo anterior implica que, para seleccionar los mejores modelos de pronóstico, se hace necesario establecer la medida de error a utilizar de antemano.

La ***Tabla 5*** establece una interpretación de la métrica.

Tabla 5

Interpretación del MAPE.

MAPE-value	Accuracy of forecast
Less than 10%	Highly Accurate Forecast
11% to 20%	Good Forecast
21% to 50%	Reasonable Forecast
More than 51%	Inaccurate Forecast

Fuente: Lewis, C., 1982.

3. WMAPE: A diferencia del MAPE, el WMAPE realiza un promedio ponderado por el peso del valor real. Por consiguiente, esta métrica minimiza los efectos de observaciones por grandes variaciones pro con poco impacto en los valores reales. El WMAPE queda definido como:

$$WMAPE = \frac{1}{\sum Y_t}\left(\sum_{t=1}^{T}\left|\widehat{Y_t} - Y_t\right|\right) \times 100$$

4. R^2: También reconocido como el coeficiente de determinación, tiene la principal virtud de que sirve como comparación al modelo base. Tal como hace referencia su nombre, el modelo base es simplemente el promedio de las observaciones. El valor de esta métrica oscila entre $-\infty$ y 1. Cuando R^2 es negativo, significa que el modelo es peor que predecir la media. Asimismo, cuando el valor es 0, se puede decir que el modelo es igual al modelo base. Finalmente, cuando el valor equivale a 1, el modelo se ajusta perfectamente a la realidad. La fórmula R^2 queda definida como:

$$R^2 = 1 - \frac{\sum_t \left(Y_t - \widehat{Y_t}\right)^2}{\sum_t (Y_t - \bar{Y})^2}$$

5. AIC: El criterio de información de Akaike estima la cantidad de información relativa perdida de un modelo. Por ende, un modelo con menos perdida de información tiene una mayor calidad. Para realizar dicha estimación, el AIC realiza un *trade-off* entre la complejidad del modelo y la bondad de ajuste. En otras palabras, busca balancear entre el sobre ajuste (al tener un modelo muy complejo) y el riesgo de sub ajustar (modelo muy simple). El AIC de un modelo se puede calcular de la siguiente manera:

$$AIC = 2k - 2ln(\hat{L})$$

donde $\hat{L}$ corresponde al valor máximo de la función de probabilidad (*likelihood*) y k a la cantidad de parámetros en el modelo. El mejor escogido es el que minimice el AIC.

6. BIC: El criterio de información bayesiano está relacionado al AIC descrito anteriormente. A diferencia del AIC, la penalización de *overfitting* no solo depende de la cantidad de parámetros, sino que también de la cantidad n de observaciones. Por consiguiente, el BIC se calcula de la siguiente manera:

$$BIC = \ln(n)k - 2ln(\hat{L})$$

2.4. Correlación

La correlación es la forma numérica en la que la estadística ha podido evaluar la relación de dos o más variables, es decir, mide la dependencia de una variable con respecto a otra variable independiente. Se dice que dos variables cuantitativas están correlacionadas cuando los valores de una varían sistemáticamente con respecto a la otra variable. En la estadística, hay dos tipos de correlaciones que son utilizados con frecuencia por parte de los investigadores.

2.4.1. Correlación de Pearson

Es un coeficiente paramétrico que infiere sus resultados a la población real, por lo que se tiene que cumplir el supuesto de normalidad en las variables. Además, es independiente de la escala de medida de las variables. El coeficiente ρ se calcula:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

donde σ_{XY} corresponde a la covarianza de las variables mientras que σ_X y σ_Y corresponde a la desviación estándar de cada variable. Los valores obtenidos ρ oscilan entre -1 y 1, donde -1 implica una correlación negativa perfecta, 0 implica una correlación nula y 1 implica una correlación positiva perfecta.

2.4.2. Correlación de Spearman

Es un coeficiente no paramétrico de la correlación de rango, es decir, calcula la dependencia estadística del ranking entre dos variables. El coeficiente ρ se calcula:

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

donde D es la diferencia entre los correspondientes estadísticos de orden de x - y. N es el número de parejas de datos.

Para poder interpretar las correlaciones, se utiliza la ***Tabla 6***, donde interpreta los diversos valores que puede tomar ρ. A grandes rasgos, se puede determinar que una correlación significativa corresponde a un ρ mayor a 0.7.

Tabla 6

Interpretación del coeficiente de correlación.

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Fuente: Hinkle D., Wiersma W. y Jurs S. *"Applied Statistics for the Behavioral Sciences"*, 2003.

2.5. Sensores

Para poder llevar a cabo la presente investigación, es necesario obtener los datos de tráfico de clientes para cada tienda. Con la tecnología actual, es posible obtener la siguiente información con la ayuda de sensores. Dichos sensores tienen la función de registrar la fecha y hora exacta en la cual una persona realiza una de las cuatro posibles actividades: Entrar a una tienda, salir de una tienda, pasar por afuera y detenerse en la vitrina de una tienda. Existen diversas tecnologías para realizar las mediciones de tráfico, las cuales se diferencian principalmente en la calidad de medición (precisión) y el costo. Los dos principales tipos de sensores son:

- Térmico: También conocido como sensor de calor. Es un dispositivo que registra una medición al identificar una fuente de calor, como el cuerpo humano. La principal desventaja de esta tecnología son el doble conteo debido a que el cliente está portando un objeto con fuente de calor como por ejemplo un café (***Figura 4***).

Figura 4

Ejemplo de un sensor térmico.

Fuente: Irisys

- 3D: Utilizan tecnología láser para captar el movimiento. Además, es el sensor con la más alta definición del mercado y adquiere medición del tráfico 5 veces más rápido que otros sensores que ofrece el mercado. (***Figura 5***).

Figura 5

Ejemplo de un sensor 3D.

Fuente: Xovis.

Una limitación de la presente investigación consiste en la precisión de los sensores instalados es de un 98%, por lo que una parte del ruido se puede asociar a la calidad del sensor. Para el propósito de esta investigación, se utilizan sensores 3D ya que estos tienen una mayor precisión en la medición que los térmicos.

3. Estado del arte

3.1. Predicción de series temporales

El presente trabajo de investigación se puede dividir en dos procesos: El pronóstico de tráfico de clientes en una tienda de *retail*; y la dotación de trabajadores con el fin de minimizar la sub y sobre dotación de trabajadores con respecto a la demanda. Con el fin de otorgar una mayor claridad al lector, se analiza el estado del arte de dichos procesos por separado. Finalmente, se presenta el estado del arte que inspiró este trabajo.

A lo largo de la historia, diversos modelos han surgido con el fin de predecir series de tiempo. Durante el siglo XIX, se realizaron los primeros intentos de estudiar series de tiempo, los cuales asumían un mundo determinístico. A principios del siglo XX, Yule (1927) establece que cada serie de tiempo puede ser considerada como un proceso estocástico. Como consecuencia, varios métodos preliminares surgieron basándose en esta simple idea. Trabajos como Slutsky, Yaglom, Walker y Yule formularon los modelos autorregresivos (AR) y medias móviles (MA). Posteriormente, Kolmogorov (1941) formula y soluciona un problema lineal. Luego, la publicación de *Time Series Analysis: Forecasting and Control* por parte de Box y Jenkins (1970) integra el conocimiento existente y desarrollara un ciclo iterativo de tres etapas para la identificación, estimación y verificación de series temporales, conocido como el enfoque de Box y Jenkins. Dicho enfoque tiene una enorme repercusión en lo que es la teoría y la practica moderna en el pronóstico de series temporales. Con el desarrollo tecnológico de las computadoras, se populariza el uso de modelos autorregresivos integrados con medias móviles (ARIMA). Cabe destacar que ARIMA es un modelo univariado. El modelo vector ARIMA (VARIMA) es una generalización multivariada del ARIMA univariado. Si bien los principales estudios fueron derivados por Quenouille (1957), no fue hasta la década de los 80 y 90 que investigadores empezaron a utilizarlo.

Otro modelo que tomo mucha fuerza en el mundo de las series temporales fue el suavizado exponencial. Este modelo se origina en la década de 1950 y 1960 con las investigaciones de Holt (1957), Brown (1959) y Winters (1960). Bajo la misma línea, Pagels (1969) expande este modelo al proveer un simple, pero útil, clasificador de tendencia y estacionalidades dependiendo de si la serie temporal corresponde a una aditiva (lineal) o multiplicativa (no lineal). En 1985, Gardner sintetiza todos los estudios de suavizado exponencial a la fecha y, además, extiende el clasificador de Pagels, incorporando tendencias amortiguadas. Dicha investigación estimula el uso de este modelo y desencadena en varios estudios adicionales. Desde entonces, han surgido varios estudios aplicados en diversos contextos tales como componentes computacionales (Gardner, 1993), pasajeros aéreos (Grubb y Masa, 2001) y planificación productiva (Miller y Liberatore, 1993).

Durante las últimas décadas, el número de nuevos modelos predictivos ha incrementado con vasta rapidez, lo cual se debe al aumento en la capacidad de procesamiento de datos por parte de los computadores. Lo anterior genera una rama en los pronósticos descrita como inteligencia artificial (IA). Modelos como redes neuronales son formulados por Hu (1964), pero debido a la falta de un

algoritmo de entrenamiento para redes con múltiples capas, el modelo es bastante limitado. En 1986, con la introducción del modelo de propagación hacia atrás (Rumelhart et al.), se comienza a desarrollar el uso de redes neuronales para predecir series temporales. Weigend et al. (1990) y Cottrell et al. (1995) estudian la estructura de la red para realizar pronósticos. Por otro lado, Tang et al. (1991), Sharda y Patil (1992) y Tang y Fishwick (1993), entre otros, reportan resultados de una serie de comparaciones entre Box-Jenkins y redes neuronales.

Otro modelo de IA es el de bosque aleatorio. El modelo fue propuesto por Ho en 1995, estableciendo que un bosque de árboles que se dividen con hiperplanos oblicuos puede ganar precisión a medida que crecen sin sufrir de sobre entrenamiento, siempre y cuando dichos bosques estén restringidos al azar. Posteriormente, Breiman (2001) elabora una extensión del algoritmo y Khoshgoftaar, Golawala y Hulse (2007) conducen experimentos para analizar el rendimiento del modelo.

Finalmente, Taylor y Letham (2017) desarrollan un modelo robusto y de fácil usabilidad, basándose en el modelo de Harvey y Peters (1990). Yenidogan et al. (2018) utilizan dicha herramienta para realizar predicciones del valor de Bitcoin, la criptomoneda más conocida actualmente.

3.2. Dotación de trabajadores

Las tiendas de *retail*, ya sean de especialidad o por departamento, se han percatado que el tráfico es fundamental en las ventas. Es por lo anterior, que gastan gran parte del presupuesto publicitario en atraer clientes mediante campañas de marketing con el propósito de poder convertir un porcentaje de dicho tráfico en ventas. Se ha encontrado evidencia de que la relación entre el tráfico y las ventas es positiva, a pesar de que la conversión (cantidad de clientes que, luego de entrar a la tienda, realizan una compra) puede en ocasiones disminuir para niveles de tráficos muy elevados. Es por lo anterior que planificar la dotación de personal respecto al tráfico puede reportar beneficios económicos para las tiendas, ya que se espera una atenuación sobre el efecto de conversión (Perdikaki et al., 2012).

Asimismo, realizar la asignación de personal de ventas en función de ventas pasadas tiene como consecuencia que se arrastren los errores cometidos en el pasado. En general las tiendas intentan mantener un *ratio* promedio de ventas y vendedores constante en el tiempo. Dicho ratio promedio no considera la variación de tráfico durante el día, específicamente durante las horas *peak* de un turno (por ejemplo, a las 3pm). Esto conlleva a que la dotación de fuerza de ventas esté sistemáticamente sub dotada en las horas de más congestión de una tienda, resultando en ventas perdidas sistemáticamente (Mani et al., 2015).

3.3. *Benchmark*

El presente trabajo de investigación busca integrar un pronóstico de series temporales robusto, utilizando como punto de referencia el trabajo realizado por Lam et al. (1998), y obtener el beneficio económico asociado a la dotación de personal, replicando el trabajo realizado por Olivares y Yung (2018).

El modelo predictivo de Lam et al. (1998) es un ARIMA (1,0,3) con los datos agregados por hora. Además, se utiliza una transformación logarítmica del tráfico y diferenciando con un desfase de una semana, cuyo objetivo es considerar estacionalidad semanal en los datos. Asimismo, componentes auto regresivos son incluidos para ajustar la auto correlación en los residuos.

4. Metodología

Para la realización de la presente investigación, se utilizan los datos de tráfico de tres clientes de Intelligenxia. El primero corresponde a una cadena de 13 tiendas de vestuario infantil. El segundo corresponde a 18 tiendas de repuestos para automóvil y, en tercer lugar, se tienen los datos de 23 tiendas de vestuario deportivo.

Posteriormente, se realiza un análisis exploratorio con el fin de identificar tendencias, estacionalidades y comportamientos de los clientes. Además, se identifican los *missing values*, los son reemplazados con un modelo de imputación de datos. Este modelo de imputación identifica un conjunto de tiendas semejantes y, a través de una regresión lineal, obtiene el valor de tráfico, para así completar el vacío. Para determinar las tiendas que poseen semejanza entre ellas, se utiliza la correlación de Pearson, la cual es la más utilizada en la academia debido a su simplicidad y facilidad de interpretación.

Luego, se realiza una serie de pronósticos entre los cuales se encuentra ARIMA, Árboles Aleatorios y la herramienta Prophet. Esta última, por ser un instrumento contingente que es utilizado actualmente por los investigadores.

Posteriormente, se utiliza el modelo desarrollado por Olivares y Yung (2018) para realizar una dotación óptima de trabajadores, usando como insumo el mejor modelo obtenido (según la métrica de comparación seleccionada) y se compara con la dotación obtenida del modelo 7, el punto de referencia. Con el fin de cuantificar la mejora en el pronóstico – y por consiguiente la dotación asignada – se calcula el ingreso esperado.

Luego, se genera un algoritmo que detecta cuando un sensor está fallando. Lo anterior se lleva a cabo a través de un control estadístico de procesos, el cual señala cuando una medición se encuentra fuera del intervalo de confianza esperado. Cuando efectivamente una medición se encuentra fuera del intervalo de confianza, el algoritmo procede a comparar el comportamiento del sensor con otros ubicados en tiendas similares (determinado por la correlación) para determinar si dicho efecto es generalizado o efectivamente el sensor en cuestión está fallando. Además, el algoritmo genera un informe detallando las fechas y tiendas de los sensores que presentan fallas, facilitando al usuario la detección de estas anomalías.

Finalmente, se establecen las principales conclusiones obtenidas del trabajo. Asimismo, se detallan las limitaciones identificadas a lo largo de la investigación en conjunto a las posibles extensiones de éste, con el fin de mejorar los algoritmos y modelos propuestos.

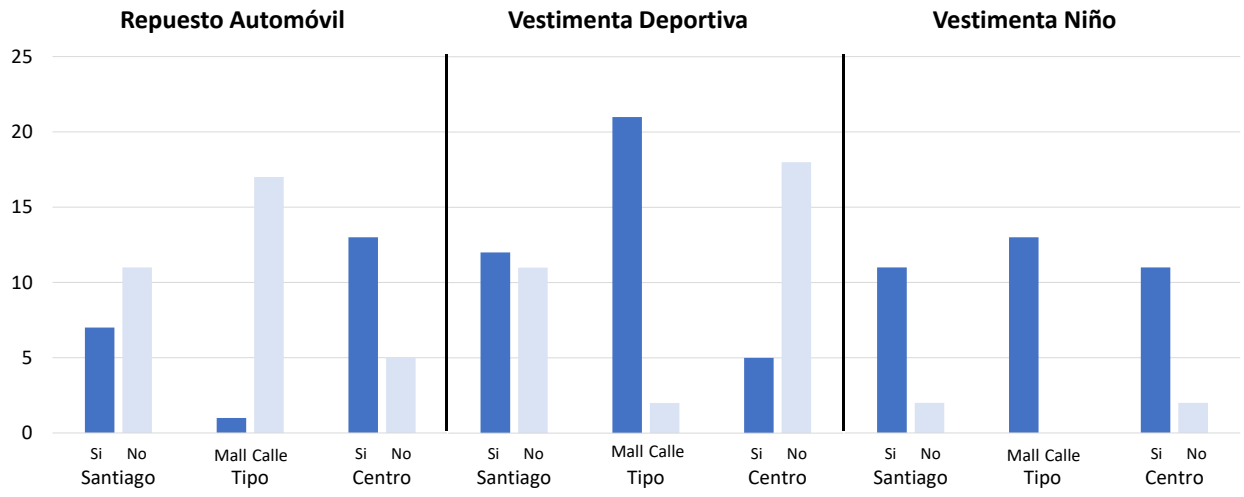
5. Obtención de los datos

Los datos son proporcionados por la empresa chilena Intelligenxia BG. Estos datos se consiguen a través de *queries* que obtienen como resultado un archivo de texto. Se realizan tres *queries*, una para cada tipo de tienda: vestimenta deportiva, repuestos de automóvil y vestimenta niño. En el trabajo de investigación se trabaja con 54 tiendas, donde 23 son tiendas de vestimenta deportiva, 18 tiendas de repuestos automóvil y 13 son tiendas de vestimenta para niños. Los datos vienen agregados por hora, donde una observación del archivo de texto significa la cantidad de personas que ingresaron en una hora, día y tienda en específico. La base de datos contempla más de 550.000 observaciones. El Anexo III muestra detalladamente los componentes de la base de datos.

Existen tres variables que caracterizan a una tienda: (1) Tienda calle o tienda *mall*, (2) tienda ubicada en el centro o fuera de él y (3) si está ubicada en Santiago o no. En la ***Figura 6*** se muestra un resumen de estas tres características y desagregado por tipo de tienda.

Figura 6

Caracterización de tiendas por rubro.

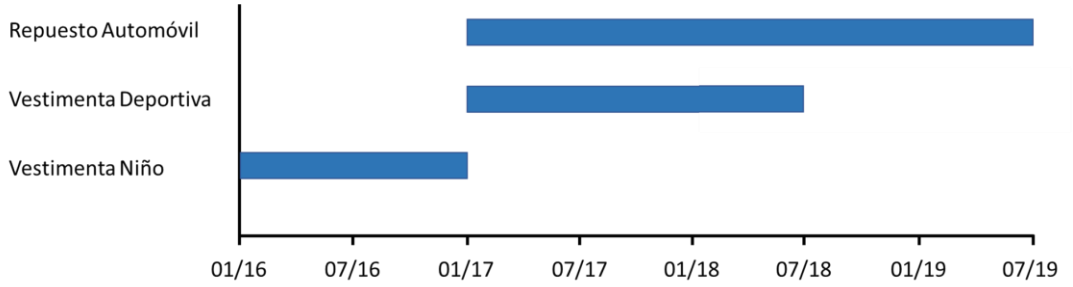


Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Además, es necesario mencionar que la *data* fue recopilada en distintos años para los distintos tipos de tiendas. Las tiendas de vestimenta deportiva contemplan desde enero 2017 hasta julio 2018, las de repuesto a partir de enero 2018 hasta julio 2019 y vestimenta infantil de enero 2016 hasta diciembre 2016 (***Figura 7***).

Figura 7

Rango temporal de los datos por tipo de tienda.



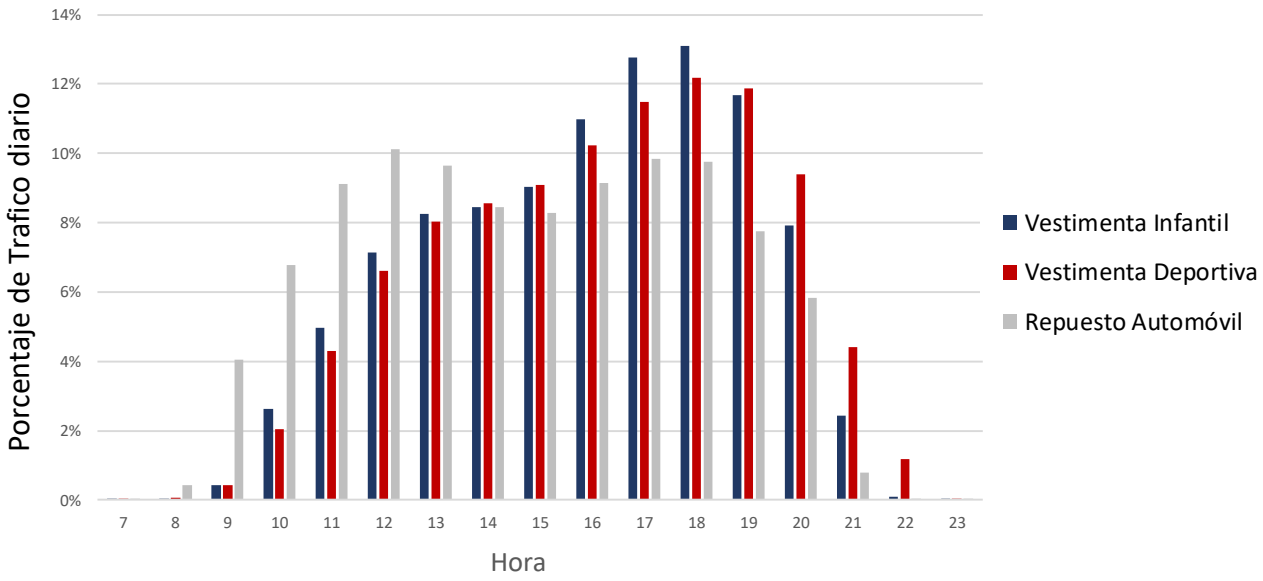
Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Debido a que los tres tipos de tiendas tienen datos en distintos momentos temporales, como consecuencia hay una mayor dificultad para obtener *insights* sobre la interacción entre estas tiendas. Para el desarrollo de la investigación, se utilizará la herramienta estadística R.

Asimismo, es importante destacar que la base de datos contiene mediciones en períodos que la tienda estaba cerrada sin trabajadores en su interior. Dichas mediciones se consideran como "clientes fantasmas" en la industria de los sensores. Si bien la proporción de mediciones fantasmas es despreciable, es necesario limpiar la base de datos, eliminando las mediciones realizadas cuando las tiendas estaban cerradas. Para identificar el horario de funcionamiento de las tiendas, se genera un gráfico con la distribución de tráfico por hora para cada tipo de tienda (***Figura 8***). Las tiendas de automóvil tienen un horario de 8 – 21, inclusive; las tiendas de vestimenta infantil de 9 – 21, mientras que vestimenta deportiva tiene un horario de 9 – 22.

Figura 8

Tráfico agregado por hora para cada tipo de tienda.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

6. Análisis exploratorio

Entender y explorar los datos es un paso fundamental antes de realizar todo tipo de análisis. El principal objetivo de la presente sección es identificar las principales características de los datos obtenidos. A través de métodos visuales y estadísticos, es posible identificar patrones, tendencias, valores atípicos entre otros.

En esta sección se identifican las distintas componentes de la serie temporal y luego se identifican los incorrectos o faltantes que tendrán que ser tratados para que éstos no afecten el rendimiento del pronóstico.

Posteriormente, se realiza el análisis exploratorio correspondiente de las series de tiempo que serán utilizadas en la presente investigación. Posteriormente, se utilizan los conocimientos obtenidos del análisis para formular los modelos de pronóstico.

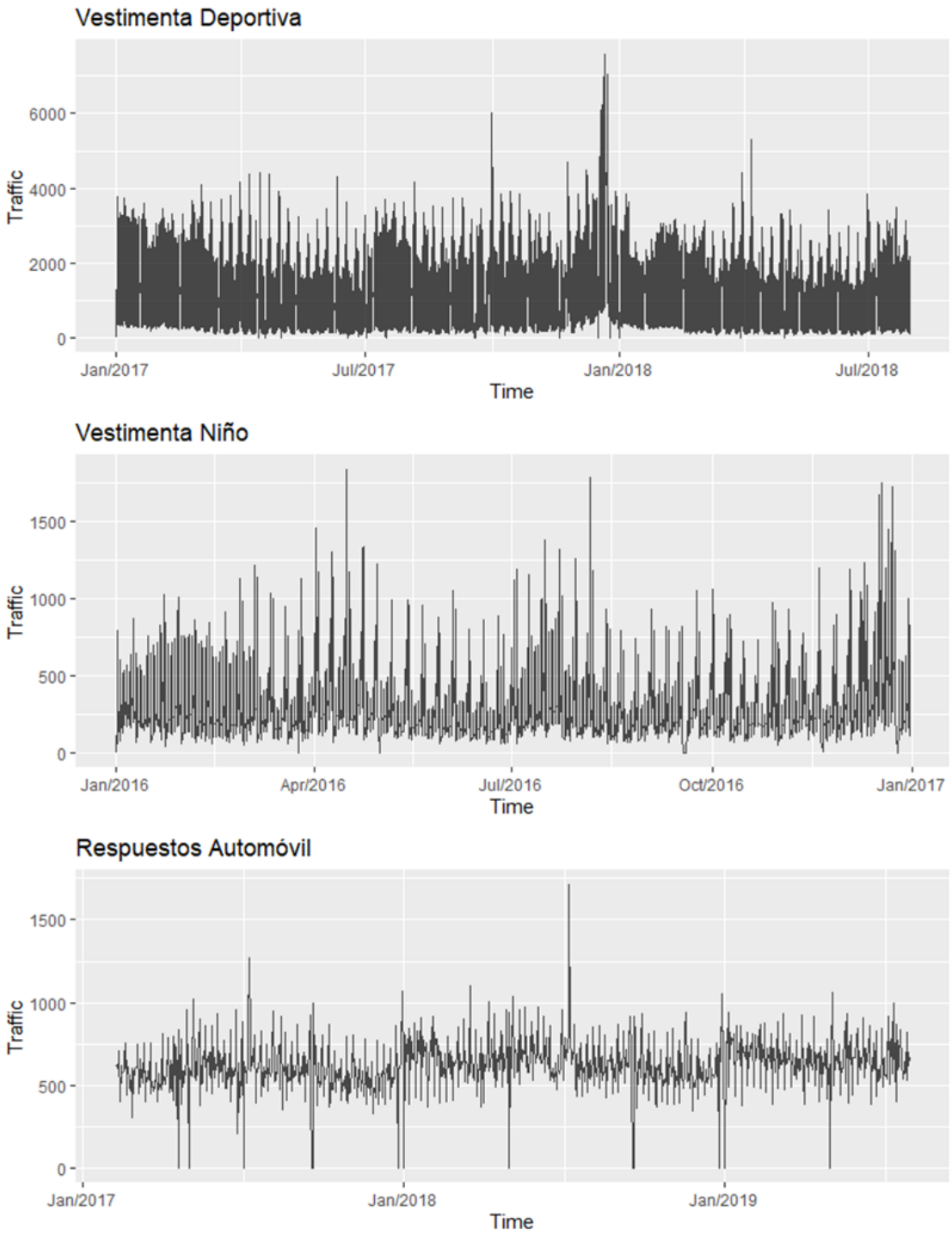
6.1. Componentes de la serie temporal

Una vez obtenido los datos crudos, es necesario realizar un análisis exploratorio de los datos para identificar los componentes de la serie de tiempo y determinar a qué tipo corresponde (aditiva, multiplicativa o mixta). Este análisis previo guía al lector a una mejor comprensión de los datos y de los componentes escogidos al momento de formular los modelos.

En primer lugar, se grafica el tráfico con la *data* original (sin preprocesamiento) de todas las tiendas para cada categoría, con el fin de obtener una idea inicial de cómo se distribuye el tráfico dependiendo del rubro (***Figura 9***).

Figura 9

Serie temporal del tráfico agregado por tipo de tienda.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

6.1.1. Estacionalidad

A partir de la ***Figura 9***, es posible determinar que la serie de tiempo se asemeja a una serie aditiva, por sobre la multiplicativa, debido a que los valores a lo largo de la serie de tiempo varían en una misma cantidad constante.

Asimismo, para las tiendas de vestimenta deportiva e infantil, se puede observar que existe una estacionalidad anual, es decir, un patrón de variación que se repite todos los años: El tráfico en enero es alto debido a las rebajas de navidad del año anterior. Luego en febrero y marzo el tráfico disminuye, pero se mantiene en niveles altos debido a las liquidaciones de fin de temporada. Posteriormente el tráfico se mantiene en niveles bajos y aumenta considerablemente en los meses de julio y agosto al ser el fin de la temporada invernal. Después, el tráfico baja en septiembre y octubre para finalmente subir y llegar a un máximo en diciembre con un *peak* en el fin de semana previo a Navidad.

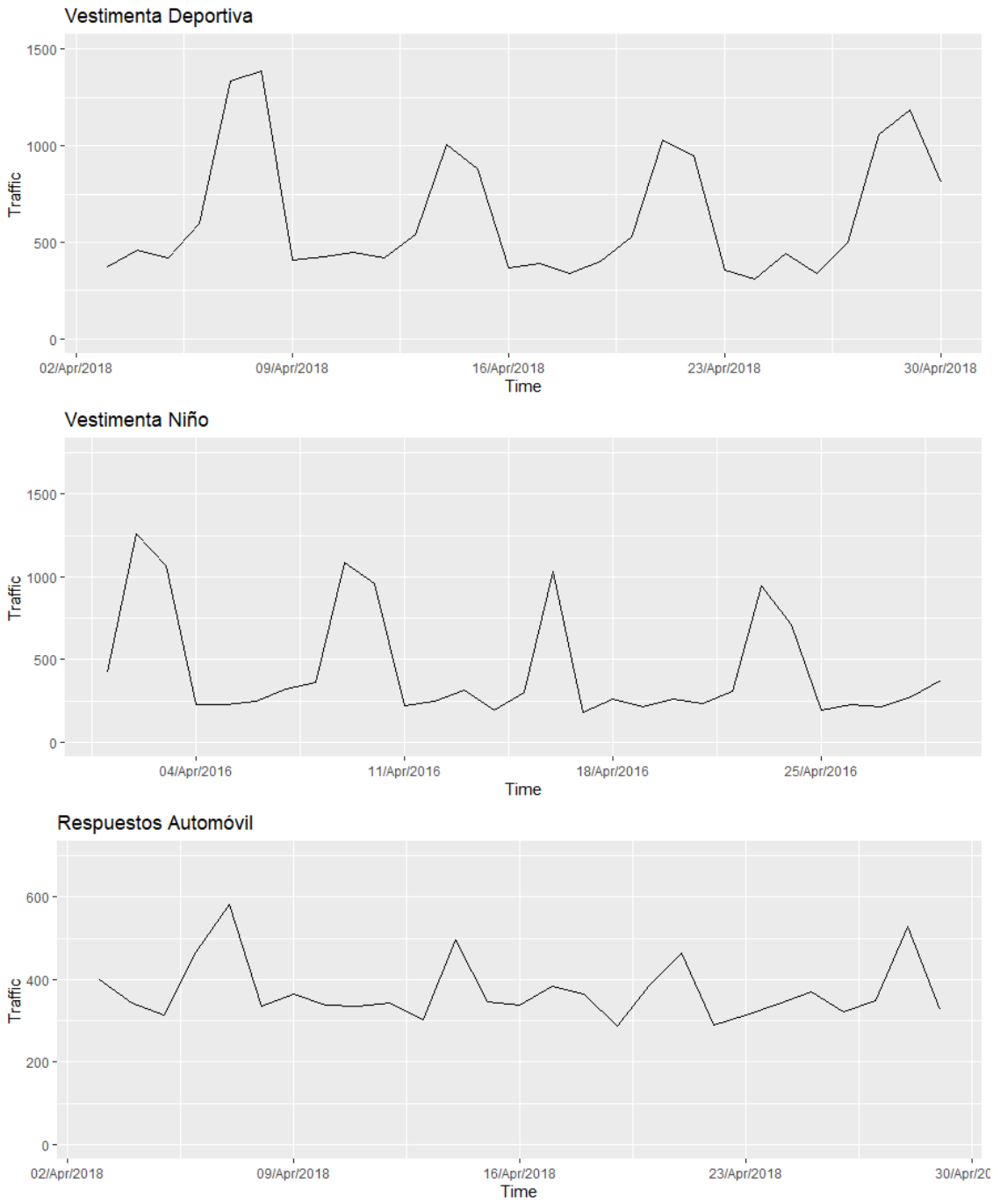
Por otro lado, las tiendas de repuesto automovilístico presentan una demanda bastante pareja a lo largo del año. Lo anterior se debe a que los repuestos de autos son demandados de manera aleatoria. Además, dicho rubro no posee liquidaciones significativas de temporada debido a que la mayoría de los productos que ofrecen son demandados durante todo el año. Sin embargo, se puede observar un leve incremento de tráfico durante el mes de febrero. En Chile, dicho mes se debe pagar el permiso de circulación para todos los vehículos motorizados, cuyo requisito es tener la revisión técnica aprobada. Es común observar personas realizando la revisión técnica a último minuto y, por ende, comprando repuestos necesarios para obtener el permiso.

Es importante mencionar que en la ***Figura 9*** se pueden observar valores nulos, los cuales corresponden a días feriados en donde la tienda se encontraba cerrada o, en una menor cantidad de casos, errores de medición por parte de los sensores. Es fácil determinar qué datos corresponden a esta última categoría ya que se posee una lista para cada tienda de todos los días que éstas estaban abiertas.

Para observar si existen otras estacionalidades con periodos menores a un año, es necesario realizar un acercamiento y graficar una ventana de tiempo menor de la Figura 9. Se grafica el tráfico de la misma tienda para el mes de mayo 2016 (***Figura 10***).

Figura 10

Ejemplo del tráfico de una tienda de cada tipo.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Se establecen como día de semana los lunes, martes, miércoles y jueves, mientras que los días de fin de semana son los viernes, sábado y domingo. A partir de la ***Figura 2*** es fácil determinar la existencia de estacionalidad semanal en el tráfico de clientes, sin importar el tipo de tienda. El tráfico de las tiendas aumenta considerablemente el fin de semana con respecto a los días de semana. Para la vestimenta deportiva, las 23 tiendas poseen mayor tráfico durante el fin de semana. De la misma manera, sólo una tienda de vestuario niño (7.7%) posee menor tráfico durante el fin de semana, mientras que, para las tiendas de repuestos de automóvil, la cifra aumenta a cuatro tiendas (22.2%). Lo anterior condice con la realidad, ya que las personas poseen mayor tiempo libre durante el fin de semana para realizar actividades de ocio (vestimenta), mientras que el efecto es menor para las tiendas de repuestos debido a que no tiene mucha relación con el ocio, sino con una necesidad.

A partir de la ***Tabla 7***, se puede observar la distribución de tráfico semanal ponderado para cada tipo de tienda. En general, el sábado es el día más concurrido con un flujo promedio de 21.3%. Para las tiendas de vestuario, el domingo es el segundo día más concurrido, mientras que para las tiendas de reparación de vehículos es el viernes. Lo anterior tiene relación con la hipótesis del párrafo anterior.

Tabla 7

Distribución del tráfico semanal promedio por tipo de tienda.

Día	Vestuario Infantil	Vestuario Deportivo	Repuesto Automóvil
Lunes	11.0%	11.9%	14.2%
Martes	10.4%	11.1%	13.3%
Miércoles	10.9%	11.1%	13.5%
Jueves	11.8%	11.1%	13.4%
Viernes	15.1%	13.9%	14.3%
Sábado	23.6%	21.7%	19.0%
Domingo	17.3%	19.2%	12.4%

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Se puede observar que las tiendas de vestuario tienen una distribución similar. El tráfico promedio durante los días de semana es de 11% y 11.3% para infantil y deportivo respectivamente, mientras que el tráfico del fin de semana es de 18.6% y 18.3% respectivamente. En cambio, las tiendas de repuesto automovilístico poseen, en promedio, un 13.6% de tráfico durante los días de semana y 15.2% los fines de semana.

Además, es interesante analizar la diferencia de la distribución semanal con respecto a las tiendas que se ubican en el centro de las ciudades en comparación con las que se ubican fuera de éste. La ***Tabla 8*** muestra dicha diferencia.

Tabla 8

Distribución del tráfico semanal promedio por tipo de tienda y ubicación.

Día	Vestuario Infantil		Vestuario Deportivo		Repuesto Automóvil	
	Centro	No Centro	Centro	No Centro	Centro	No Centro
Lunes	11.6%	10.9%	12.9%	11.7%	14.7%	14.0%
Martes	10.7%	10.4%	11.9%	10.9%	13.7%	13.1%
Miércoles	10.5%	10.9%	11.8%	11.0%	13.7%	13.4%
Jueves	11.8%	11.8%	12.0%	10.8%	13.8%	13.2%
Viernes	14.7%	15.1%	14.7%	13.6%	14.6%	14.2%
Sábado	22.3%	23.7%	20.2%	22.1%	18.9%	19.0%
Domingo	18.4%	17.2%	16.5%	19.9%	10.7%	13.1%

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

En la figura anterior, se puede observar que el promedio de los días de semana (lunes a jueves) para las tiendas ubicadas en el centro es de 11.1%, 12.1% y 14.0% para las tiendas vestuario infantil, vestuario deportivo y repuesto automóvil respectivamente. Estos valores son mayores que las tiendas ubicadas fuera de este, con 11.0%, 11.1% y 13.4% respectivamente. En promedio, existe una diferencia de 1.1 puntos porcentuales. En consecuencia, se establece como hipótesis que las tiendas ubicadas en el centro obtienen mayor flujo durante los días de semana debido a que ahí se ubican las principales fuentes de trabajo y, por consiguiente, una mayor cercanía con el potencial cliente.

Como primer contra argumento, se puede señalar que en el centro existen mayor proporción de tiendas calle (6) en relación con las tiendas *mall* (5), por lo que el efecto no surge por estar ubicado en el centro sino por el *layout* de una tienda calle en comparación con la tienda *mall*. Las tiendas calle son de fácil acceso y más pequeñas, por lo que se necesita menos tiempo libre.

Para comprobar o refutar el argumento anterior se genera la ***Tabla 9***, donde se muestra la diferencia de distribución entre las tiendas deportivas en un *mall* y calle ubicadas en el centro

Tabla 9

Distribución del tráfico semanal promedio de las tiendas de vestuario deportivo por tipo de tienda y ubicación.

Día	Centro y Calle	Centro y Mall
Lunes	13.2%	11.6%
Martes	12.2%	10.7%
Miércoles	12.1%	10.6%
Jueves	12.3%	10.8%
Viernes	15.0%	13.7%
Sábado	19.7%	22.3%
Domingo	15.6%	20.3%

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

En primer lugar, es importante mencionar que la comparación por tipo de tienda se realiza solamente entre las tiendas de vestuario deportivo debido a que estas son las únicas que poseen establecimientos en el centro y que sean de ambos tipos. A partir de la ***Tabla 9***, se puede observar que las tiendas calle ubicadas en el centro poseen un tráfico promedio de 12.4% durante los días de semana, en comparación con un 10.9% para las tiendas *mall*, es decir, 1.5 puntos porcentuales de diferencia.

Dado lo anterior, se rechaza la hipótesis establecida anteriormente debido a que el aumento de tráfico en los días de semana no se debe al estar ubicada en el centro, sino que se debe a la combinación de estar en el centro (cercano a los potenciales clientes durante la semana) y ser tienda calle (conveniente para el poco tiempo libre que se tiene durante la hora de almuerzo o posterior al horario de oficina).

Para entender de mejor manera los efectos de las variables en el tráfico durante los días de semana, se procede a realizar una regresión lineal. $TráficoPromedio_i$, que representa el tráfico promedio de la tienda i durante los días de semana, es la variable dependiente. Las variables independientes son tres variables *dummy*: $Centro_i$, donde es 1 si la tienda se encuentra en el centro de la ciudad y 0 si no; $Calle_i$, donde es 1 si es una tienda calle y 0 si es una tienda *mall*; $Santiago_i$, donde es 1 si la tienda se encuentra en Santiago y 0 si no. La ***Tabla 10*** se muestra los resultados obtenidos.

Tabla 10

Regresión lineal de la distribución del tráfico semanal.

Variable	A1	A2	A3
Intercept	0.117 (***)	0.116 (***)	0.117 (***)
Centro	0.011 (***)	0.012 (***)	0.009 (**)
Calle		0.015 (***)	-0.005
Santiago		-0.001	
Centro Calle			0.014 (*)
Number of variables	2	4	4
Degrees of freedom	54	52	52
Residual error	0.011	0.009	0.009
R2	0.435	0.622	0.674
F-Statistic	32.28	21.93	27.56

P-values: *** (<0.1%), ** (<1%), * (<5%)

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

La ***Tabla 10*** proporciona información interesante a destacar. En primer lugar, se crea el primer modelo (A1) el cual cuantifica el efecto de las tiendas ubicadas en el centro para los días de semana. Se obtiene que una tienda ubicada en el centro aumenta, en promedio, su tráfico en día de semana en 1.1 puntos porcentuales. Este efecto concuerda con lo descubierto anteriormente en la Figura 4. Luego, se genera el segundo modelo (A2), el cual consiste en agregar al modelo anterior las variables binarias $Centro_i$ y $Santiago_i$. Se obtiene que dichas son significativas mientras que la variable Santiago no lo es. Además, se obtiene un R^2 de 0.622. En tercer lugar, se crea el último modelo (A3) el cual incluye las variables $Centro_i$, $Calle_i$ y su interacción entre ellas $Centro_i * Calle_i$. Se obtiene que una tienda (ya sea *mall* o calle) ubicada en el centro posee 0.9 puntos porcentuales más de tráfico, en comparación con una tienda ubicada fuera de éste. También, en el caso que dicha tienda sea de tipo calle, el efecto es más pronunciado, con un aumento de 2.3 puntos porcentuales en el tráfico en comparación con una tienda calle fuera del centro. El resultado anterior concuerda con la teoría debido a que los trabajadores, que usualmente se concentran en el centro de la ciudad y tienen poco tiempo libre, prefieren (1) tiendas cercanas, es decir ubicadas en el centro y, (2) que sean tiendas calle por su facilidad de acceso.

Finalmente, a partir de la ***Figura 10***, se puede pensar que existe una estacionalidad mensual donde el tráfico disminuye a medida que pasan los días del mes y luego vuelve a subir al comienzo del mes subsiguiente. El argumento anterior puede tener sentido debido que normalmente las personas reciben el sueldo a fin de mes. Según Pagel (2016), cuando las personas reciben su sueldo, se sienten más ricos y por ende sienten que pueden consumir una porción del sueldo inmediatamente, aún cuando las personas saben de antemano que recibirán dicho ingreso. Esta respuesta psicológica ha sido comprobada en Inglaterra, donde el 33% de los encuestados respondieron que gastan más en los primeros días posteriores de recibir el sueldo (Macleod, 2014). Dicho efecto psicológico es conocido por algunas empresas. Tiendas de supermercados, por ejemplo, suelen subir los precios a comienzos del mes para maximizar las utilidades provenientes de los clientes que son pagados a fin de mes.

Para determinar si existe el efecto mencionado anteriormente se procede a calcular, para todas las tiendas, la demanda porcentual de cada día con respecto a la demanda semanal. Posteriormente, se eliminaron todas las semanas que poseen algún día feriado debido a que afectan directamente la métrica calculada. Además, se eliminaron los meses de enero y diciembre debido a que estos tienen comportamientos distintos al resto del año y alteran los resultados. Por ejemplo, en diciembre el tráfico aumenta a medida que avanza el mes, hasta llegar a un máximo los días previos de Navidad, mientras que a comienzos de enero se puede observar las rebajas post Navidad sumado a los clientes que van a cambiar los regalos.

Para comprobar el efecto estacional en los primeros días de cada mes, se propone realizar una prueba de posición de dos muestras, la cual comprueba si las medias de dos poblaciones, que distribuyen normalmente, son semejantes. La primera muestra son los primeros cuatro días del mes mientras que la segunda muestra serán los días restantes. Sin embargo, de manera preliminar se necesita comprobar que las medias entre los meses sean constantes, con el fin de evitar efectos de estacionalidad. Se obtiene que la distribución de tráfico se comporta de la misma manera para los meses considerados, con excepción de febrero, el cual presenta variaciones significativas. Lo anterior se puede deber a que el mes de febrero es cuando la mayor cantidad de trabajadores están de vacaciones, lo cual es posible altere dichas distribuciones. El resto de los meses no presenta variaciones estadísticamente significativas y, en el caso que hubiera, se pueden considerar mínimas para el cálculo a continuación. En promedio, las tiendas de vestuario infantil presentaron una diferencia entre las medias mensuales de 0.90% con una varianza de 0.071%. Con estos datos, se obtiene un T estadístico de 0.92 que no es suficiente para establecer una diferencia entre meses. Para las tiendas de vestuario deportivo se obtiene un T estadístico de 1.02, mientras que en las tiendas de repuesto automóvil es de 0.96.

Posteriormente, se procede a comprobar la estacionalidad mensual. Se establece como hipótesis nula (H_0) que la media para cada día de la semana es la misma para los primeros días del mes $\{1,2,3,4\}$ que para el resto de los días (5 o más). Dado que la hipótesis es por día de la semana, se realizará el mismo procedimiento 7 veces para cada tipo de tienda. La ***Tabla 11*** muestra un resumen de los resultados obtenidos.

Tabla 11

Impacto de los primeros días del mes en el tráfico.

	Vestimenta Infantil	Vestimenta Deportiva	Repuesto Automóvil
Lunes	-0.538	1.011 (***)	-0.608 (**)
Martes	1.211 (**)	-0.416 (**)	-0.471 (**)
Miércoles	-0.421	0.615 (*)	-0.187
Jueves	-0.843 (*)	-0.038	0.053
Viernes	0.487	0.290	-0.062
Sábado	2.174 (***)	0.859 (*)	0.522 (**)
Domingo	3.494 (***)	2.735 (***)	0.016

P-values: *** (<0.1%), ** (<1%), * (<5%*)

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

A partir de la figura anterior, se destacan dos características. En primer lugar, el tráfico en las tiendas de repuesto automovilístico no se ve afectado por el número de día del mes, en comparación con las tiendas de vestimenta. Lo anterior se puede deber a que los repuestos automovilísticos son necesarios y corresponden a un problema que debe ser resuelto a la brevedad, en comparación con las tiendas de vestimenta que es una compra que puede esperar y no es urgente. En segundo lugar, se puede observar un efecto positivo y significativo durante los días sábado y domingo para las tiendas de vestimenta, y para el sábado en las tiendas de repuesto de automóvil.

En conclusión, se puede determinar que existe un efecto de estacionalidad mensual en el tráfico de clientes, específicamente durante los primeros días de cada mes. Dicha estacionalidad tiene distinto grado de magnitud, la cual depende del tipo de tienda y su flexibilidad de compra por parte del usuario.

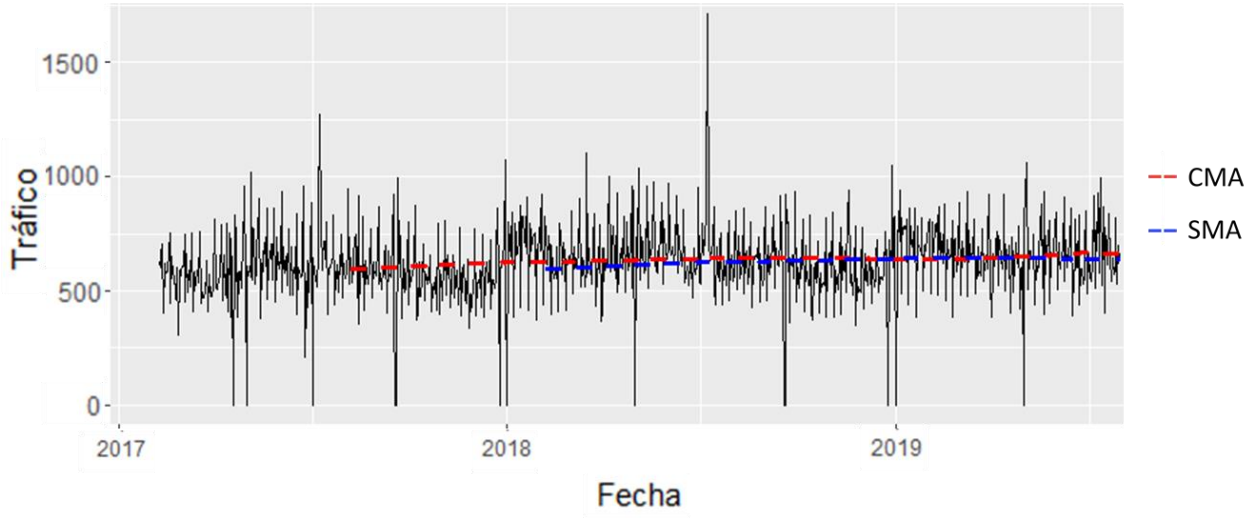
6.1.2. Tendencia

Otro componente elemental en las series de tiempo es la tendencia secular o tendencia a largo plazo. Dicho componente caracteriza el patrón gradual y consistente de las variaciones de una serie de tiempo. Usualmente, las variaciones están relacionadas con factores a largo plazo que afectan el crecimiento o reducción de la serie.

Para obtener el valor de la tendencia en el transcurso de los años, la literatura recomienda dos indicadores: Media Móvil Centrado (CMA) y Media Móvil Simple (SMA). La ***Figura 11*** representa la serie de tiempo para una tienda de vestimenta deportiva con las dos medias móviles descritas anteriormente.

Figura 11

Serie de tiempo con media móvil simple y centrada para una tienda de repuesto automovilístico.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Al utilizar medias móviles para visualizar la tendencia, la única decisión que se debe tomar es la anchura w de ventana a utilizar. Para medir correctamente la tendencia, se debe considerar una ventana de anchura equivalente a la estacionalidad más grande. En esta investigación se considera $w = 365$, correspondiente a la estacionalidad anual.

Para ver la tendencia global a lo largo de las 54 tiendas estudiadas, se procede a calcular el CMA para dichas tiendas. Cabe resaltar que se decide utilizar CMA sobre SMA para calcular la tendencia, debido a que el CMA utiliza los datos adyacentes, lo cual implica que tanto la serie original como la CMA van a tener en los mismos momentos los valores máximos y mínimos.

Como resultado, se obtiene un aumento en la tendencia del tráfico de un 7.8% en 18 meses para las tiendas de repuesto automovilístico. Por otro lado, se obtiene una disminución del tráfico en un 5.5% para las tiendas de vestuario deportivo. Finalmente, no se puede calcular la tendencia para las tiendas de vestuario infantil debido a que solamente se posee un año de datos.

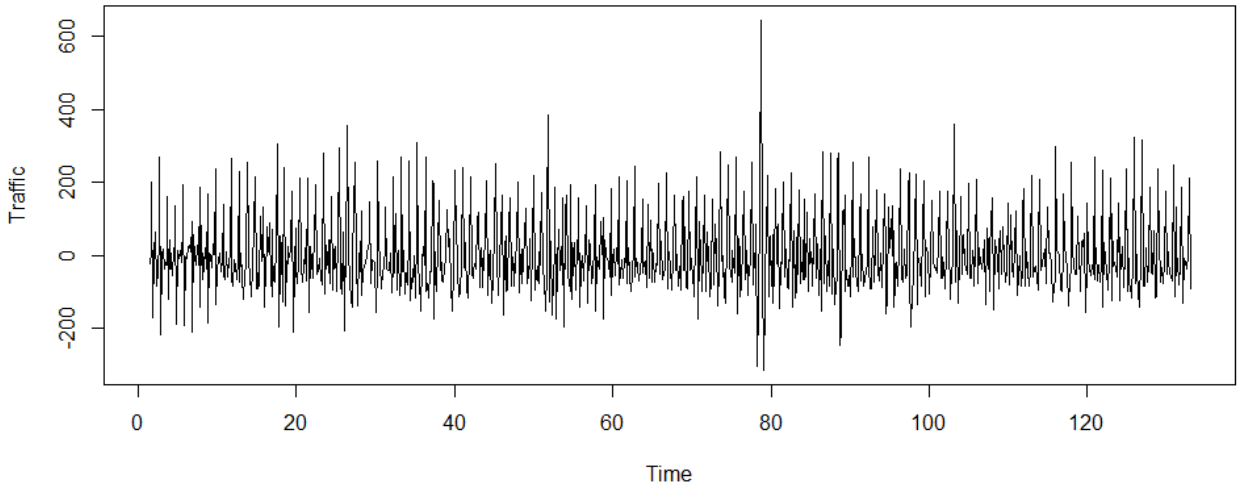
La baja en el tráfico deportivo es bastante sorprendente y se le pueden asociar tres causas principales. En primer lugar, la industria del *retail* en Chile es bastante competitiva, sobre todo en las tiendas de vestuario. Lo anterior implica que la disminución de tráfico se puede deber a la pérdida de ventajas competitivas en comparación con los principales competidores del país, lo cual se traduce en una pérdida de tráfico hacia la competencia. En segundo lugar, las ventas online han canibalizado lentamente el tráfico de los clientes de tiendas físicas. El cliente chileno se arriesga cada vez más a comprar online. Según un estudio de la Cámara de Comercio de Santiago (CCS), el 4.8% de las ventas del *retail* fueron de manera online en 2017 y se proyecta un 9.3% para 2020. Lo anterior implica, con crecimiento de la industria nulo, una fuga de ventas físicas a online que explica en cierta medida la disminución de tráfico. En tercer lugar, la economía chilena ha sufrido un estancamiento importante a partir de 2018 con la guerra comercial en Estados Unidos y China, lo cual se puede trasladar en un cambio de comportamiento del consumidor chileno por el miedo a una eventual recesión económica.

6.1.3. Ruido

Luego de descomponer la serie de tiempo aditiva en la tendencia y estacionalidades, se obtiene la diferencia entre estos componentes y la serie original, el cual corresponde al ruido asociado a una serie. Para ejemplificar, se grafica la serie temporal de ruido para una tienda de repuestos automovilístico (***Figura 12***).

Figura 12

Componente de ruido para la tienda de repuesto automovilístico.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

6.2. Clasificación de las series temporales

Como ya fue mencionado anteriormente, las series de tiempo se pueden clasificar en dos grupos: series estacionarias y no-estacionarias. Es importante clasificar las series de tiempo del presente trabajo de investigación, debido a que las series estacionarias son más fáciles de predecir, ya que la media y varianza es constante a lo largo de toda la serie de tiempo. La gran mayoría de los modelos predictivos utilizados actualmente usan como supuesto que la serie temporal es estacionaria.

Para determinar a qué grupo pertenece cada una de las 54 series temporales, se procede a realizar la prueba de Dickey-Fuller aumentada (ADF). Como resultado, se obtiene que la mayoría de las tiendas son series temporales estacionarias con solo siete tiendas siendo no estacionarias. De las siete tiendas, cinco pertenecen a los comercios de vestuario infantil y dos a los de vestuario deportivo. Además, de estas siete tiendas, solo una tiene un p-valor de 0.29, mientras que los seis restantes tienen un p-valor menor a 0.1 (umbral es 0.05).

Es necesario clarificar que no se van a modificar las siete tiendas cuya serie no es estacionaria, debido a dos motivos: (1) No se alinea con el foco de la investigación y (2) solo una tienda presenta un efecto significativo, por lo que, en términos de resultados del trabajo, éste se puede ignorar.

6.3. *Missing values* e imputación de datos

Debido a una serie de motivos, la *data* cruda posee errores donde el valor del tráfico no fue medido. Dichos errores normalmente se deben a una remodelación del local (que puede prolongarse por algunos días o incluso semanas), la falla del sensor o del Wi-Fi que transmite el valor del sensor al servidor de datos. Estas situaciones, al ser extraordinarias y no representan la normalidad, aumentan el ruido de los pronósticos. Con el fin de mejorar la predicción final, es necesario realizar una imputación de datos, es decir, sustituir valores no informados por otros generados con un algoritmo.

Debido a que se poseen series de tiempo de varias tiendas, para sustituir los valores faltantes, se utilizará como referencia la tienda con mayor semejanza a esta. El algoritmo para la imputación de datos faltantes se puede describir en 3 pasos:

1. Computar correlación entre la tienda por arreglar y el resto de las tiendas (del mismo tipo).

2. Realizar una regresión lineal con la(s) tienda(s) de mayor correlación.

3. Insertar valores imputados dado los valores de la(s) tienda(s) con mayor correlación.

En el paso 2, se debe definir la cantidad k de tiendas a considerar en la regresión lineal. Un valor bajo de k puede significar un sub-ajuste en la predicción mientras que un valor muy alto puede significar un sobreajuste. Para obtener el valor k ideal para cada tipo de tienda y comprobar la eficacia del algoritmo, en primer lugar, se procede a realizar una prueba con valores ya conocidos que serán considerados como valores faltantes para luego realizar la imputación de datos con el algoritmo descrito anteriormente. Se consideran 12 tiendas con una semana de *data* faltante. Asimismo, para evitar sesgo de la estacionalidad, cada semana faltante será en un mes distinto. Se repite la prueba anterior para los tres tipos de tienda. Los resultados se muestran en la ***Tabla 12***.

Tabla 12

Precisión (MAPE) de la imputación de datos desagregado por tipo de tienda.

	MAPE
Vestimenta Infantil	15.83%
Respuestos Automóvil	13.05%
Vestimenta Deportiva	16.61%

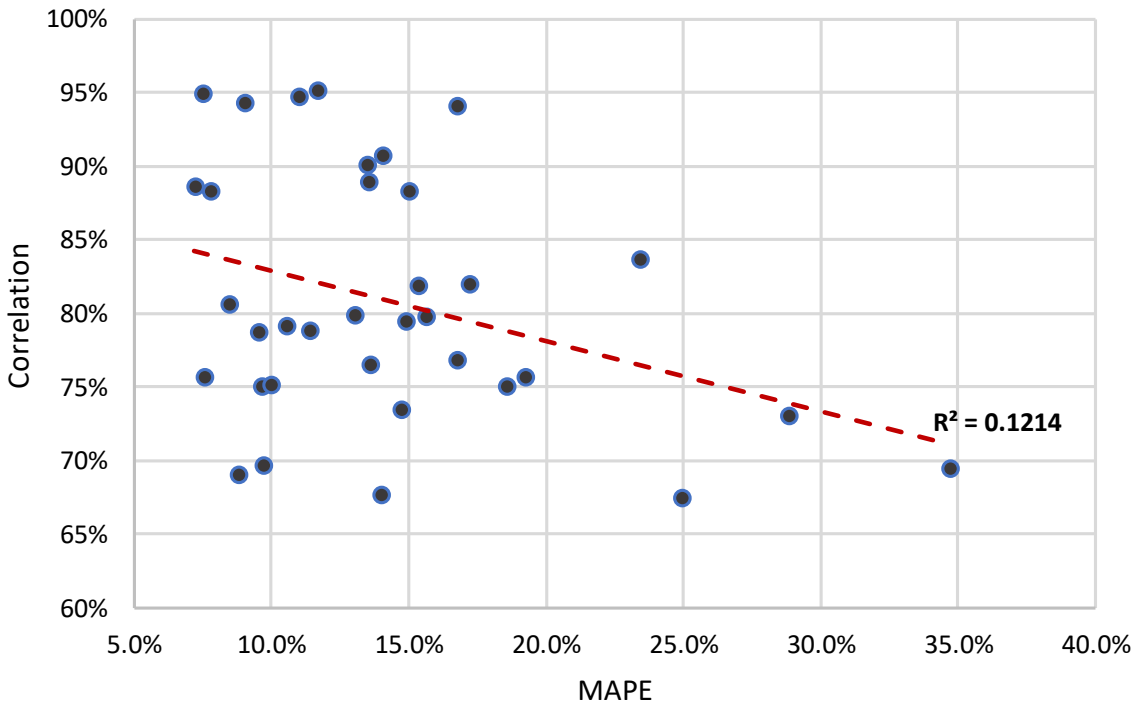
Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Se obtiene un MAPE promedio de 15.16% en la imputación de datos, con las tiendas de repuesto obteniendo una mayor precisión (13.05%) y la vestimenta deportiva la peor (16.61%). A partir de la ***Tabla 12***, se puede comentar que la imputación de datos realizada tiene una buena precisión. Asimismo, se utilizó solamente una tienda ($k = 1$) similar para realizar los pronósticos ya que una mayor cantidad de tiendas no mejoraba la predicción.

Además, se puede plantear la hipótesis que una tienda con *data* faltante que posee una tienda similar con mayor correlación tiende a obtener mayor precisión en las predicciones, es decir, un menor MAPE. Para demostrar la hipótesis mencionada anteriormente, se procede a graficar la correlación y el MAPE de las 36 semanas imputadas, obteniéndose la ***Figura 13***.

Figura 13

Relación MAPE y correlación de la imputación de datos.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

A partir de la ***Figura 13***, se puede aceptar la hipótesis planteada anteriormente. Si bien la correlación entre las tiendas implica en la calidad de la precisión, el R cuadrado obtenido (0.12) es bastante bajo, lo cual implica que existen otras variables intrínsecas de cada tienda que influyen en la calidad de las predicciones.

Con las pruebas realizadas satisfactoriamente, se procede a realizar la imputación de datos para las tiendas con *data* faltante. Para identificar los días con *data* faltante se propone la siguiente metodología que contiene cuatro pasos:

1. Para cada tienda, filtrar las fechas con tráfico mayor a cero con el fin de obtener una lista de días con tráfico faltante.

2. Eliminar las fechas que son considerados feriados (renunciables e irrenunciables).

3. Eliminar las fechas donde la suma de todas las tiendas pertenecientes al mismo tipo sea mayor a cero.

4. Imputar tráfico a las fechas restantes.

Es interesante realizar dos observaciones con respecto al algoritmo descrito anteriormente. En primer lugar, se han eliminado las fechas con clientes fantasmas, obteniéndose 0 tráfico para dichas fechas. Para considerar una tienda fantasma, el tráfico diario debe ser menor a 10. En segundo lugar, el paso tres considera la suma de las tiendas para determinar si todas las tiendas se encuentran cerradas, y por ende no hay *data* faltante, o bien existen tiendas abiertas por lo que un tráfico nulo implica que nos enfrentamos a un *missing value*. Con el algoritmo anterior, se puede determinar cuándo para una determinada fecha, la tienda posee *data* faltante.

En consecuencia, se realiza la detección de la *data* faltante para cada una de las 54 tiendas a estudiar. Se detectan 91 *missing values*: 52 para las tiendas de repuesto automovilístico, 6 y 33 para las tiendas de vestuario infantil y deportivo respectivamente. Lo anterior equivale en promedio a 2 por tienda. Posteriormente, se realiza la imputación de datos para los 91 faltantes.

7. Modelos de predicción

A continuación, se detallan nueve modelos con distintas características, complejidad y fundamentos matemáticos, que fueron mencionados y explicados en el marco conceptual. Posteriormente, se comparan y analizan los resultados obtenidos, con el objetivo de obtener el modelo con mayor poder predictivo.

7.1. Presentación de modelos: aplicación y resultados

Los modelos son desarrollados utilizando R, un lenguaje de programación con enfoque al análisis estadístico. Actualmente, es uno de los programas más manejados en la investigación y cada vez es más utilizado por las empresas debido a que soporta herramientas de aprendizaje automático, minería de datos y matemáticas financieras, temáticas contingentes en la actualidad profesional.

Para medir la eficiencia de los modelos presentados a continuación, se realizarán pronósticos que serán posteriormente comparados con los datos originales. Para las tiendas de vestuario deportivo y repuestos automovilístico, se separa la serie de tiempo en dos subconjuntos: conjunto de entrenamiento, que son los datos con lo que se construye el modelo, y conjunto de validación, el cual sirve para validar el modelo y prevenir errores de ajuste. Los modelos serán comparados por su precisión predictiva en el conjunto de validación, el cual en otras palabras significa realizar un pronóstico fuera de la muestra utilizada para entrenar el modelo.

Para las tiendas de repuesto automovilístico, el conjunto de validación incluye los meses de junio y julio 2019. Por otro lado, para las tiendas de vestuario deportivo, dicho conjunto incluye los mismos meses, pero para el año 2018. Finalmente, las tiendas de vestuario infantil tienen como conjunto de validación octubre 2016. Para las tiendas de vestuario infantil, se realiza un pronóstico dentro de la muestra (*in-sample*) para los modelos 3 - 9, es decir, el mes de octubre 2016 está incluido en el conjunto de entrenamiento. Además, es necesario destacar que el modelo para las tiendas de vestuario infantil no incluye información de la estacionalidad anual debido a que el conjunto de entrenamiento es menor de 12 meses.

Teniendo en consideración la diferencia de precisión entre un pronóstico dentro de la muestra y fuera de ella, se realizan ambos pronósticos para las tiendas de vestuario deportivo y repuesto automovilístico. La finalidad de esto es poder cuantificar la diferencia de precisión y realizar comparaciones entre la tienda de vestuario infantil con las restantes.

En resumen, la ***Tabla 13*** detalla el tipo de pronóstico a realizar por rubro:

Tabla 13

Tipos de pronósticos realizados por tipo de tienda.

Rubro	Detalle
Vestuario Infantil	dentro
Vestuario Deportivo	dentro y fuera
Repuesto Automóvil	dentro y fuera

Fuente: Creación propia.

Para definir la calidad de los modelos, se utiliza la métrica MAPE, debido a ser la ésta es la más utilizada por la academia. Para calcular el MAPE promedio de las 54 tiendas, se consideran los pronósticos fuera de la muestra, ya que de esta forma se logra un resultado más certero. En caso de que un tipo de tienda no posea pronóstico fuera de la muestra, se considerarán los pronósticos dentro de ella.

Los primeros dos modelos corresponden a pronósticos básicos de baja complejidad, los cuales sorpresivamente son los modelos utilizados por la gran mayoría de la industria del *retail*. El tercer, cuarto y quinto, corresponden a modelos ARIMA. El tercero corresponde al modelo formulado por Lam y Vandenbosch (1996), mientras que el cuarto es el modelo desarrollado por Olivares y Yung (2018). El quinto corresponde a un modelo de creación propia, realizando una pequeña reformulación del modelo anterior. El sexto, séptimo y octavo modelo corresponden a formulaciones que utilizan la herramienta Prophet. Finalmente, el noveno modelo es formulado utilizando árboles aleatorios. Los últimos cuatro modelos son formulados por el autor.

Para evaluar de manera tangible los modelos generados en el presente informe, es necesario tener un punto de comparación (*benchmark* en inglés). Dado lo anterior, es necesario mencionar que el presente trabajo surge de la motivación de mejorar los pronósticos realizados por Olivares y Yung (2018). En consecuencia, la sección 7.2, se presenta un análisis comparativo entre los modelos formulados por parte del autor y el modelo de Olivares y Yung (2018).

Es necesario mencionar que los datos utilizados por Olivares y Yung (2018) son los mismos de las tiendas de vestimenta infantil. Lo anterior implica que el modelo escogido dependerá del rendimiento general, mas también del rendimiento específico de las tiendas de vestuario infantil.

Además, y a modo ejemplificativo, se muestra una figura para cada modelo el cual corresponde al pronóstico de una tienda de repuestos automovilísticos ubicada en Santiago. Se utiliza una tienda de repuesto automovilístico por sobre vestuario infantil debido a que el pronóstico del primero corresponde a uno realizado fuera de la muestra, por lo que es más riguroso. También, el segundo modelo aplica solamente a los pronósticos fuera de la muestra. Asimismo, cabe resaltar que es solo una ejemplificación y el rendimiento de una tienda en específico no refleja el rendimiento del modelo en su totalidad. En la sección 8, se evalúa el impacto económico, a través de la dotación de trabajadores, del mejor modelo predictivo escogido con respecto al modelo de Olivares y Yung (2018).

7.1.1. *Naive* I

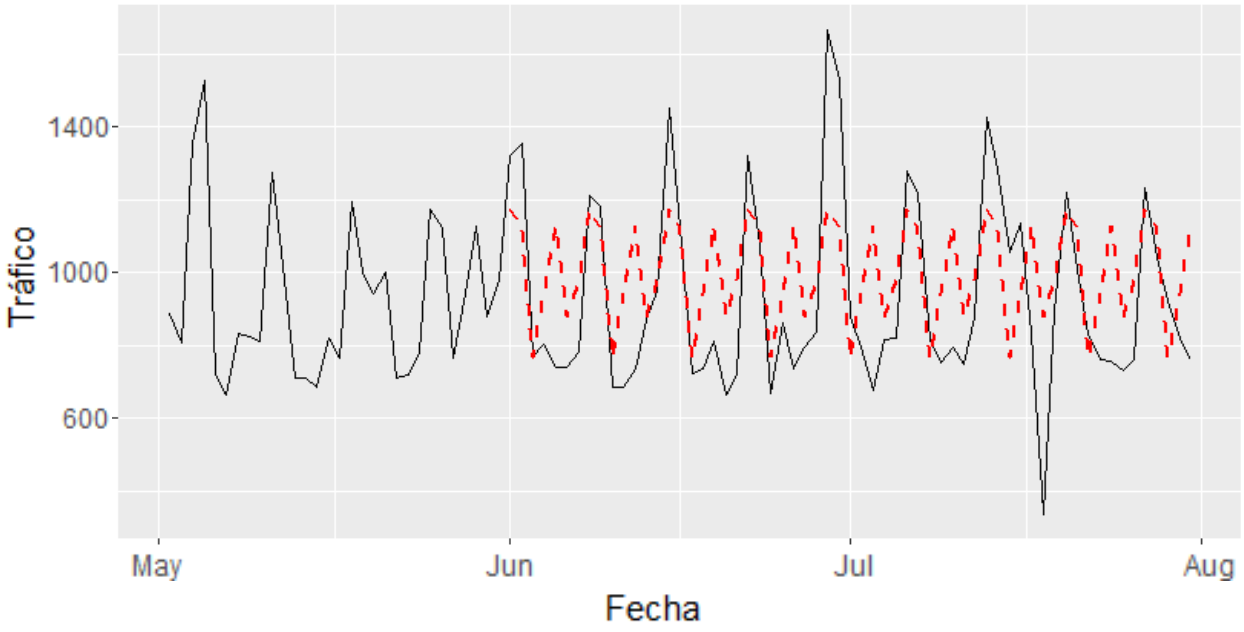
Si bien es tentador implementar modelos complejos y sofisticados, es necesario encontrar el modelo con mayor poder predictivo y que, al mismo tiempo, sean fáciles de implementar. Dado lo anterior, el presente modelo se caracteriza por su simpleza y actualmente es uno de los más ocupados en la industria del *retail* al momento de pronosticar tráfico de clientes en tienda como también ventas.

El pronóstico *Naive* para series estacionales simplemente establece cada pronóstico equivalente al último valor observado de la temporada pasada. En este caso, como la estacionalidad semanal es la que impacta de mayor manera la serie de tiempo, se considera el valor del mismo día en la semana pasada para realizar el pronóstico. La lógica de este pronóstico es que la información reciente es la más útil para poder predecir el futuro, basándose en que la industria es dinámica y cambia todos los años.

Utilizando el modelo descrito en el párrafo anterior, se realiza el pronóstico de las 54 tiendas en cuestión. Para la tienda utilizada como ejemplificación, se obtiene la ***Figura 14***. La línea negra corresponde a los datos originales y la línea roja el pronóstico.

Figura 14

Pronóstico Naive I para una tienda de repuesto automovilístico. MAPE = 26.20%.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

En promedio, se obtiene un MAPE de 26.20%. Los resultados por tipo de tienda se presentan en la ***Tabla 14***.

Tabla 14

Resultados pronóstico Naive I por tipo de tienda.

Tienda	MAPE
Vestimenta Infantil	30.37%
Repuestos Automóvil	17.29%
Vestimenta Deportiva	30.55%

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

A partir de los resultados, se pueden obtener dos conclusiones. En primer lugar, el modelo, aún siendo bastante simple, puede lograr buenas predicciones, por lo que no es extraño que actualmente la industria lo siga utilizando con el objetivo de predecir el futuro flujo de clientes. En segundo lugar, se pueden identificar varias limitaciones del modelo, principalmente la pequeña ventana de predicción.

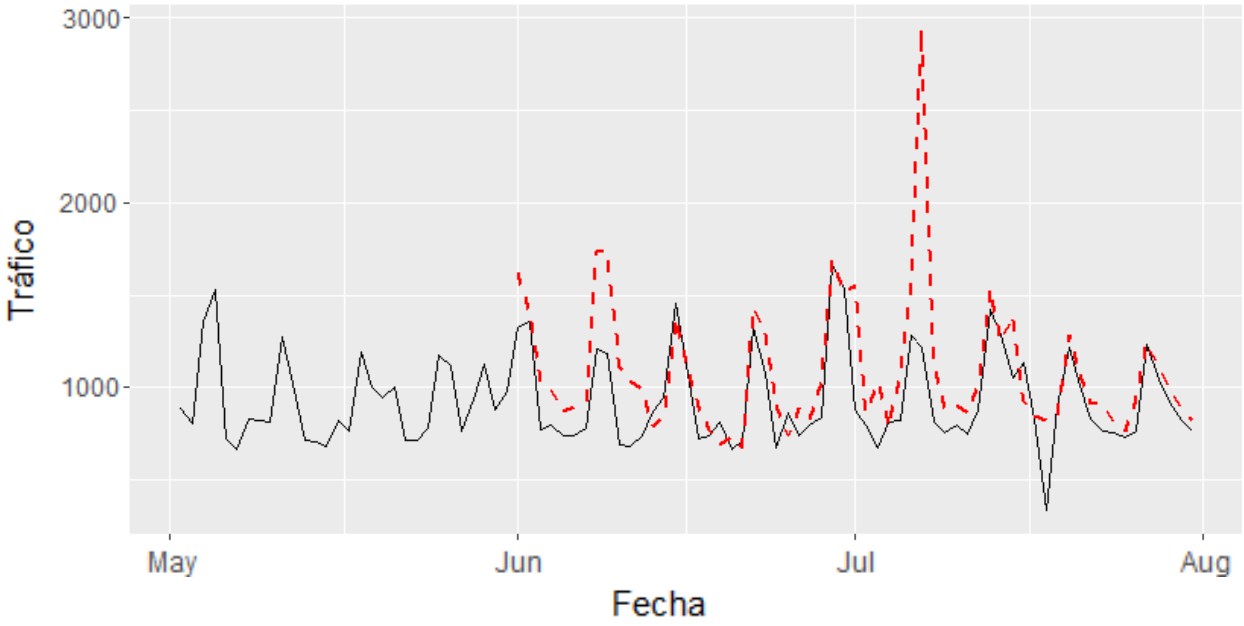
Es imprescindible mencionar que el modelo anterior no sirve para pronosticar meses debido a que no considera la estacionalidad anual que fue demostrada en la sección anterior. Además, el modelo no considera días especiales como los feriados, por lo que no se logra capturar la variación de tráfico para dichos días.

7.1.2. *Naive* II

Se propone modificar el modelo anterior, agregando la estacionalidad anual. A diferencia del modelo anterior, el cual usaba un *lag* – 7 para predecir el tráfico futuro, el presente modelo utiliza un *lag* – 364, es decir, 52 semanas. Es importante que el *lag* sea múltiplo de 7 para poder comparar los mismos días de semana (por ejemplo, lunes con lunes), con el fin de no tener sesgo debido a la estacionalidad semanal. Además, debido las nuevas características del modelo, es necesario tener tiendas con más de un año de *data*, por lo que no podrá ser utilizado en las tiendas de vestuario infantil. Utilizando el modelo descrito, se realiza el pronóstico de las 41 tiendas. Para la tienda utilizada como ejemplificación, se obtiene el siguiente pronóstico (***Figura 15***)

Figura 15

Pronóstico Naive II para una tienda de repuesto automovilístico. MAPE = 26.43%.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

En promedio, se obtiene un MAPE de 26.43%. Los resultados por tipo de tienda se presentan en la ***Tabla 15.***

Tabla 15

Resultados pronóstico Naive II por tipo de tienda.

Tienda	MAPE
Vestimenta Infantil	N/A
Repuestos Automóvil	19.85%
Vestimenta Deportiva	31.58%

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Se puede identificar una serie de connotaciones del nuevo modelo. En primer lugar, es necesario destacar que la complejidad del presente modelo (*Naive* II) es mayor que el modelo previo (*Naive* I). Sin embargo, a partir de las Figuras 4.1 y 4.3, se puede observar que el modelo *Naive* II se sobre ajusta considerablemente, incrementando la posibilidad de generar un mal pronóstico. Por ejemplo, de la ***Figura 15***, al observar detenidamente, se puede detectar un *outlier* para el domingo 7 de julio 2019. En consecuencia, el modelo sobreestima el tráfico en más del doble del real.

Lo anterior se debe a que el modelo utiliza el valor observado hace 364 días para generar la predicción, es decir, 8 de julio 2018. El día mencionado corresponde a una promoción anual llamada "el día del repuesto", el cual ofrecía descuentos del 45% en todos los productos. Dicha promoción también se realizó en julio 2017 pero no se realizó en 2019.

Asimismo, las tiendas de vestimenta deportiva sufren del mismo problema. Se detecta un *outlier* para el domingo 1 de julio 2018 donde el modelo pronostica que todas las tiendas se encontrarían cerradas. En este caso, el modelo considera las elecciones primarias realizadas el 2 de julio 2017, donde todas las tiendas deben estar cerradas por ley.

Dado estos dos casos de *outliers* y con el objetivo de determinar el verdadero impacto predictivo de ambos modelos, se procede a eliminar dicho día del pronóstico. Se obtiene que la precisión en términos del MAPE mejora en 0.32 y 0.11 puntos porcentuales para las tiendas de repuestos y vestuario deportivo respectivamente. Debido a que los *outliers* corresponden a fechas especiales que pueden afectar negativamente a ambos modelos de la misma manera, es posible eliminar estas y determinar que el modelo *Naive* II es más preciso que *Naive* I.

Los dos modelos *Naive* formulados anteriormente son conocidos por su baja complejidad. Los siguientes modelos propuestos en el presente informe poseen una complejidad considerablemente mayor, por lo cual debiesen tener una mayor precisión. Asimismo, dichos modelos son los más utilizados actualmente por la academia y son utilizados por las empresas que están migrando a una transformación digital en la industria del *retail*.

7.1.3. ARIMA I

El primer modelo ARIMA surge de la investigación realizada por Lam y Vandenbosch (1996). Primero, usan una transformación logarítmica en la variable de tráfico y_t con el fin de introducir una serie estacionaria con varianza constante. Luego, aplican una diferenciación con un lag de una semana a la variable logarítmica transformada anteriormente, con el fin de incorporar la estacionalidad semanal. Finalmente, componentes autorregresivos (AR) y de medias móviles (MA) son incorporadas para ajustar las autocorrelaciones en los residuos. El modelo se puede describir a través de las siguientes ecuaciones:

$$x_t = ln(y_t)$$

$$w_t = y_t - y_{t-7}$$

$$w_t = \mu + \frac{\theta(B)}{\varphi(B)}\varepsilon_t$$

donde y_t: tráfico de clientes,

x_t: transformación logarítmica del tráfico de clientes,

w_t: La variable x_t diferenciada con un lag de 7 días,

μ: La media de la variable diferenciada w,

$\varphi(B)$: El componente autorregresivo,

$\theta(B)$: El componente de media móvil,

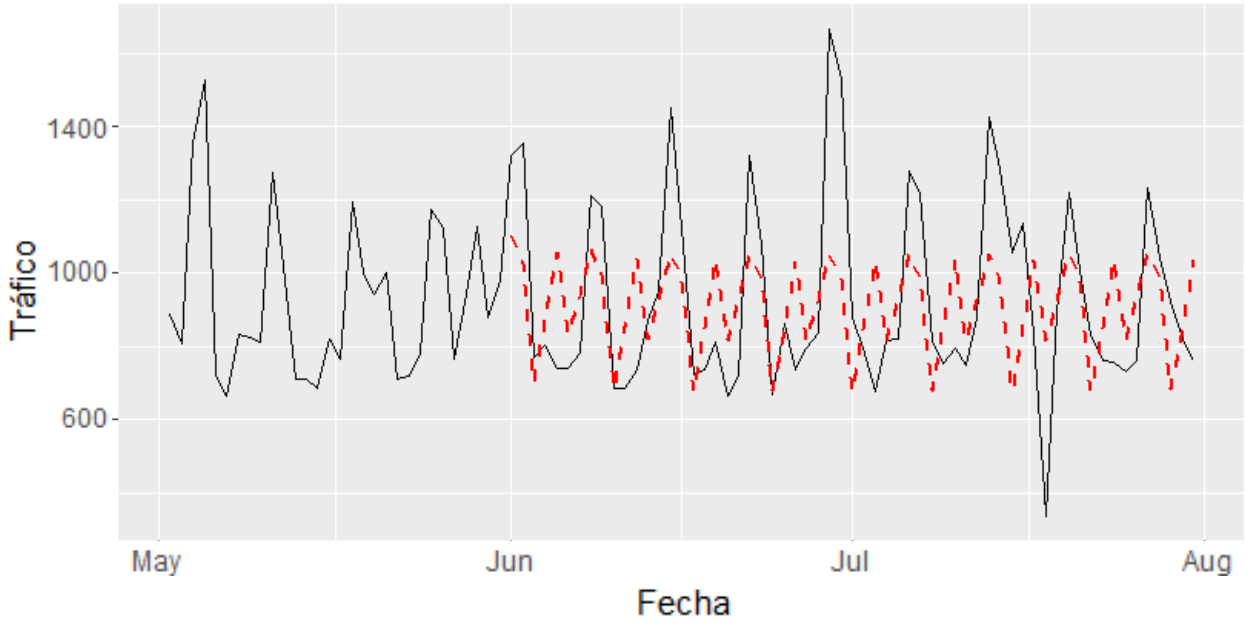
ε_t: Termino de error (residual) con media nula y varianza constante.

Actualmente, varias investigaciones de pronóstico se basan en el estudio realizado por Lam y Vandenbosch (1996).

Utilizando el modelo descrito, se realiza un modelo ARIMA (1,0,3) para el pronóstico de las 54 tiendas. Para la tienda utilizada como ejemplificación, se obtiene el siguiente pronóstico (***Figura 16***). El pronóstico (línea roja) fue obtenido con datos fuera de la muestra.

Figura 16

Pronóstico ARIMA I para una tienda de repuesto automovilístico. MAPE = 20.09%.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

A partir de la figura anterior, se puede observar que el modelo es bastante simple y prácticamente no hay diferencia de pronóstico entre las semanas. Lo anterior se debe a la falta de variables explicativas que generan variabilidad en el tráfico de clientes, como por ejemplo los días festivos. El MAPE promedio de las 54 tiendas es de 28.48%. Los resultados por tipo de tienda se presentan en la ***Tabla 16***.

Tabla 16

Resultados pronóstico ARIMA I por tipo de tienda.

Tienda	MAPE	
	Dentro	Fuera
Vestimenta Infantil	**36.12%**	N/A
Repuestos Automóvil	20.17%	**16.47%**
Vestimenta Deportiva	40.79%	**33.55%**

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

7.1.4. ARIMA II

A partir de las limitaciones anteriores, Olivares y Yung (2018) replica el modelo anterior, pero con dos modificaciones. En primer lugar, agrega una variable binaria $Holiday_t$, donde equivale a 1 cuando el día es feriado y 0 en caso contrario. En segundo lugar, los académicos realizaron modelos ARIMA con distintas especificaciones de parámetros (p, d, q), resultando el modelo ARIMA (2,0,1) el mejor dado las métricas de AIC y BIC. Es por esto, que se utiliza la especificación anterior en vez del ARIMA (1,0,3) realizado por Lam y Vandenbosch (1996). Lo anterior se debe porque el modelo ARIMA (2,0,1).

El modelo anterior queda especificado de la siguiente manera:

$$x_t = ln(y_t)$$

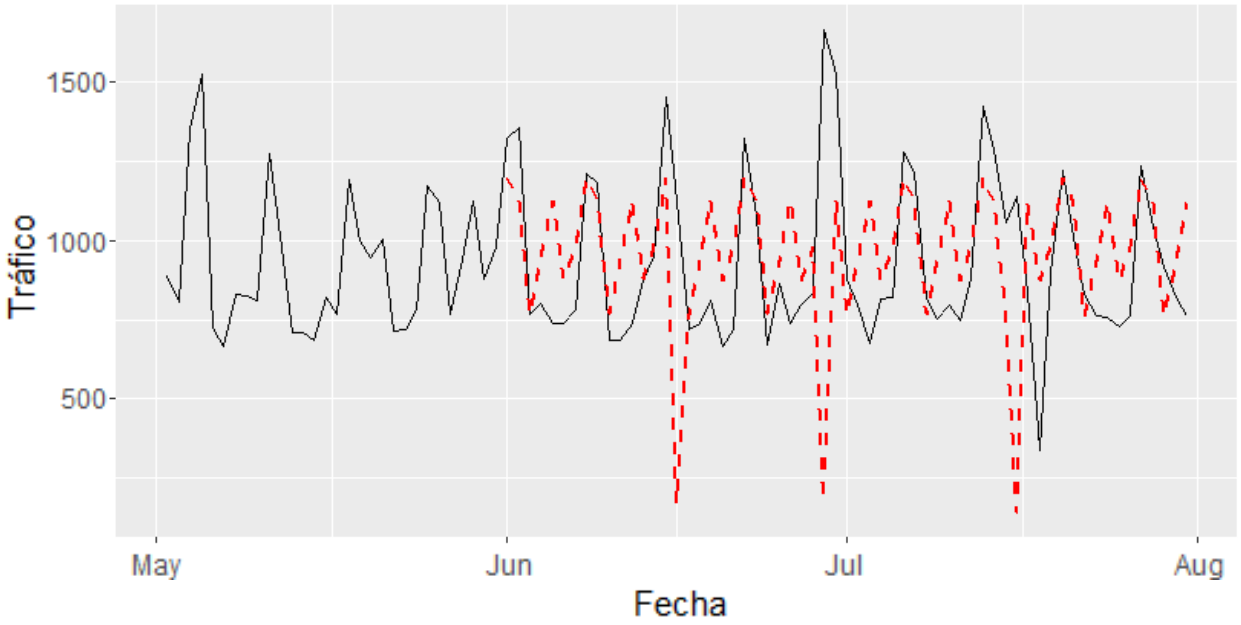
$$w_t = y_t - y_{t-7}$$

$$w_t = \mu + \alpha\, Holiday_t + \frac{\theta(B)}{\varphi(B)}\varepsilon_t$$

Utilizando el modelo descrito, se realiza un modelo ARIMA (2,0,1) para el pronóstico de las 54 tiendas. Para la tienda utilizada como ejemplificación, se obtiene el siguiente pronóstico (***Figura 17***). El pronóstico (línea roja) fue obtenido con datos fuera de la muestra.

Figura 17

Pronóstico ARIMA II para una tienda de repuesto automovilístico. MAPE = 25.33%.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

A partir de la figura anterior, se puede observar que el modelo considera el efecto de feriados, el cual ocurre en tres ocasiones entre junio y julio 2019. Además, es posible señalar un efecto negativo en el tráfico cuando el día corresponde a un día feriado. En específico, para la tienda de la ***Figura 17***, el efecto de un día festivo implica una disminución del tráfico en 839 según el modelo estimado.

Sin embargo, no todos los feriados son iguales por naturaleza. En Chile existen principalmente dos tipos de feriados: irrenunciables y no irrenunciables. El primero corresponde a feriados que la actividad económica permanece cerrada para todo comercio no esencial. El segundo corresponde a feriados religiosos los cuales las tiendas permanecen abiertas y perciben un aumento considerable en el tráfico de clientes. En la ***Figura 17***, los tres feriados corresponden a feriados no irrenunciables, por lo que el tráfico debiese aumentar y no disminuir.

Bajo este contexto, el modelo descrito por Olivares y Yung (2018) presenta un MAPE promedio de 28.65% para las 54 tiendas. La ***Tabla 17*** presenta los resultados por tipo de tienda.

Tabla 17

Resultados pronóstico ARIMA II por tipo de tienda.

Tienda	MAPE	
	Dentro	Fuera
Vestimenta Infantil	**28.24%**	N/A
Repuestos Automóvil	20.06%	**20.54%**
Vestimenta Deportiva	31.49%	**35.24%**

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Finalmente, el hecho que el modelo ARIMA ($p = 2, q = 1$) sea el mejor, implica que los modelos con más parámetros están sobre ajustados a la muestra y no se ajustan correctamente a los datos del conjunto de prueba (Box et al., 2008).

7.1.5. ARIMA III

La finalidad del tercer modelo ARIMA es proponer un modelo similar a los anteriores, modificando este último al separar la variable Holiday en dos variables binarias: $Irrenunciable_t$ y $Feriado_t$. Tal como dicen sus nombres, la primera corresponde a los días con feriados irrenunciables mientras que la segunda corresponde a un festivo normal. Se espera que el efecto del primero sea negativo mientras que el segundo sea positivo. Se utiliza la misma cantidad de parámetros establecido en el modelo de Olivares y Yung (2018), es decir, ARIMA (p = 2, q = 1).

El modelo queda descrito de la siguiente manera:

$$x_t = ln(y_t)$$

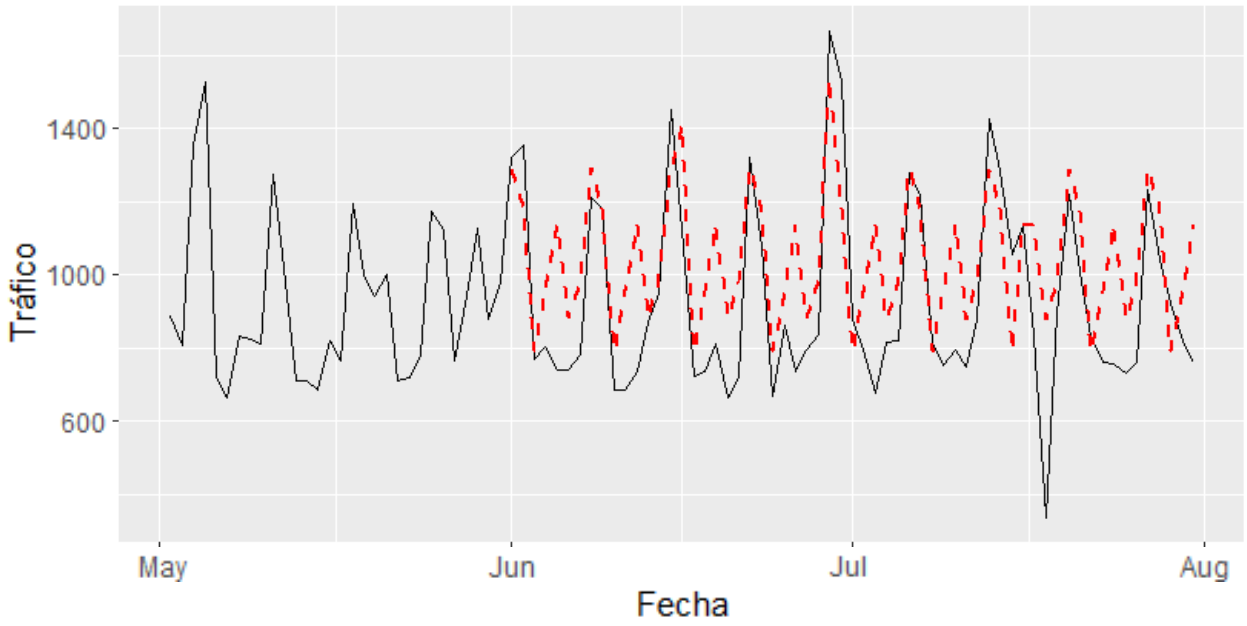
$$w_t = y_t - y_{t-7}$$

$$w_t = \mu + \alpha\, Irrenunciable_t + \beta\, Feriados_t + \frac{\theta(B)}{\varphi(B)}\varepsilon_t$$

Utilizando el modelo descrito, se realiza un modelo ARIMA (2,0,1) para el pronóstico de las 54 tiendas. Para la tienda utilizada como ejemplificación, se obtiene el siguiente pronóstico (***Figura 18***). El pronóstico (línea roja) fue obtenido con datos fuera de la muestra.

Figura 18

Pronóstico ARIMA III para una tienda de repuesto automovilístico. MAPE = 21.87%.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

A partir de la ***Figura 18***, es posible notar el efecto de considerar los dos tipos de días festivos (irrenunciables y feriado religioso). Específicamente para el ejemplo graficado, se tiene que el tráfico aumenta en promedio en 195 personas para los feriados no irrenunciables. Lo anterior tiene sentido con respecto a la hipótesis establecida anteriormente. Además, el modelo considera efectivamente los días irrenunciables como días que las tiendas se encuentran cerradas, es decir, con tráfico nulo. Lo anterior significa el tráfico en un día irrenunciable se reduce en el tráfico promedio de los días sin feriados, específicamente para el ejemplo anterior, en 1028.

Bajo este contexto, el modelo modificado presenta un MAPE promedio de 25.42% para las 54 tiendas. La ***Tabla 18*** presenta los resultados por tipo de tienda.

Tabla 18

Resultados pronóstico ARIMA III por tipo de tienda.

Tienda	MAPE	
	Dentro	Fuera
Vestimenta Infantil	**22.96%**	N/A
Repuestos Automóvil	17.56%	**17.23%**
Vestimenta Deportiva	23.71%	**33.21%**

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Con la finalidad de contextualizar al lector, se estiman los tres modelos descritos anteriormente para la tienda de repuesto automovilístico (***Tabla 19***). Se puede observar que los signos de $Irrenunciable$ y $Feriado$ tienen los signos esperados. Además, el tercer modelo es sustancialmente superior a los dos anteriores, con un AIC bastante menor.

Tabla 19

Distintas especificaciones del modelo ARIMA para una tienda de repuesto automovilístico.

	AR(1) MA(3)	AR(2) MA(1)	AR(2) MA(1)
μ	0.001	-0.001	-0.0004
	(0.001)	(0.040)	(0.013)
Holiday		-1.913	
		(0.114)	
Irrenunciable			-6.836
			(0.042)
Feriado			0.174
			(0.028)
ARMA			
	0.841 ***	0.198	0.640 ***
	(0.024)	(1.536)	(0.175)
		-0.003	-0.029
		(0.135)	(0.069)
	-0.834 ***	-0.110	-0.348 *
	(0.040)	(1.536)	(0.171)
	-0.103 *		
	(0.048)		
	-0.063		
	(0.038)		
LL	-1314.0	-1220.1	98.3
AIC	2639.9	2452.3	-182.7
BIC	2668.3	2480.6	-149.6

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

7.1.6. Prophet I

A continuación, se utilizará la herramienta creada por Facebook, Prophet. El primer modelo por utilizar incluye las tres estacionalidades comprobadas en la sección anterior: Anual, mensual y semanal. La herramienta cuenta con varias funcionalidades que son fáciles de utilizar. En primer lugar, el modelo a *priori* ya considera estacionalidad semanal y anual, por lo que no se debe agregar ningún comando para incorporar dichas estacionalidades. Además, la herramienta ofrece la opción de agregar una nueva estacionalidad donde el usuario puede definir el periodo de dicha estacionalidad. Por ejemplo, se puede incorporar la estacionalidad mensual definiendo un periodo de 30.5 días. Sin embargo, los meses no son equivalentes entre sí, por lo que el periodo de la estacionalidad mensual no es constante. Además, el efecto mensual que se busca capturar es el aumento de tráfico durante los primeros cuatro días del mes. Por consiguiente, se considera la estacionalidad mensual en el modelo utilizando un regresor *dummy*, el cual toma el valor 1 si la fecha pertenece a los primeros cuatro días del mes y 0 en caso contrario.

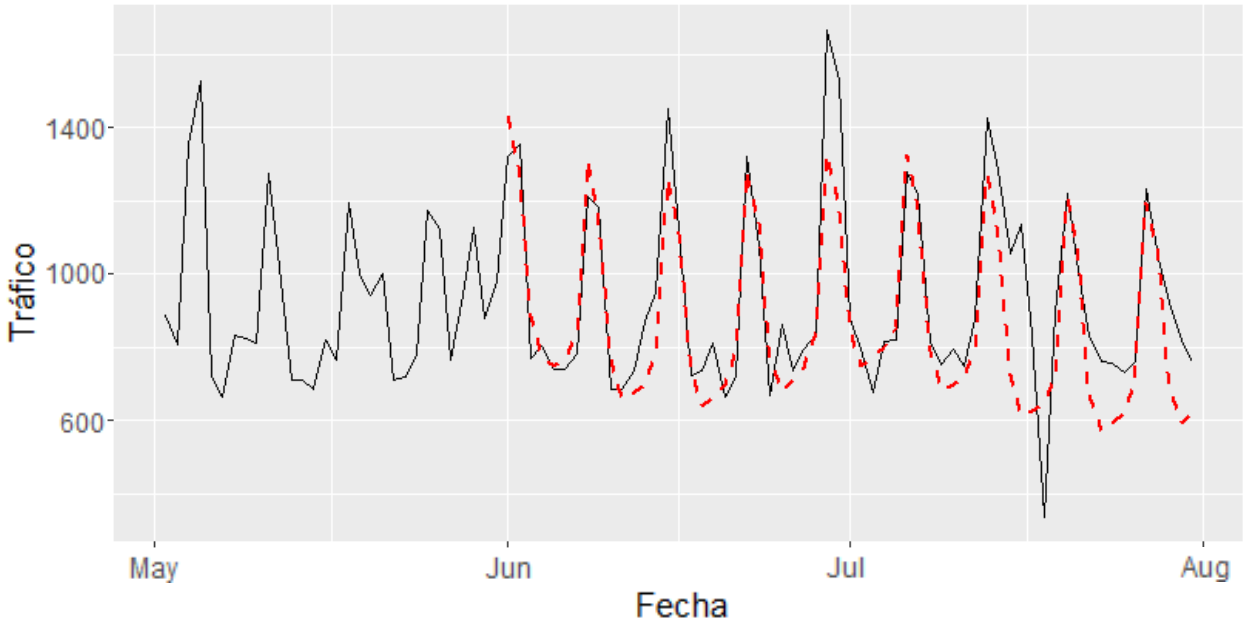
En conclusión, el modelo queda definido de la siguiente manera:

$$y_t = Trend_t + YearSeasonality_t + WeekSeasonality_t + FirstDays_t + \varepsilon_t$$

Utilizando el modelo descrito, se realiza el pronóstico de las 54 tiendas. Para la tienda utilizada como ejemplificación, se obtiene el siguiente pronóstico (***Figura 19***). El pronóstico (línea roja) fue obtenido con datos fuera de la muestra.

Figura 19

Pronóstico Prophet I para una tienda de repuesto automovilístico. MAPE = 12.47%.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Los resultados obtenidos del primer modelo de Prophet son robustos. En promedio, el MAPE correspondiente a las 54 tiendas es de 20.17. Los resultados por tipo de tienda se presentan en la ***Tabla 20***. En negrita se puede observar los datos utilizados para calcular la efectividad del modelo.

Tabla 20

Resultados pronóstico Prophet I por tipo de tienda.

Tienda	MAPE	
	Dentro	Fuera
Vestimenta Infantil	**16.81%**	N/A
Repuestos Automóvil	10.77%	**12.79%**
Vestimenta Deportiva	17.58%	**27.85%**

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Si bien, el modelo es bastante preciso, aún hay mejoras que se pueden llevar a cabo para obtener pronósticos más robustos. A partir de la ***Figura 19***, se puede observar que el modelo no se ajusta a valores extremos como los feriados. Es por esto, que se propone un nuevo modelo idéntico al actual, pero considerando los feriados.

7.1.7. Prophet II

Incluir días especiales en Prophet es directo. Se debe crear una tabla con todas las fechas especiales en el pasado y el futuro que se busca pronosticar. Además, como la mayor cantidad de los días especiales se repiten anualmente, es importante asignar un nombre constante para cada feriado a lo largo de los años, debido a que Prophet considera cada feriado como una variable binaria. Con una correcta asignación de nombres, el modelo puede capturar el efecto de cada festivo por si solo. Dado lo anterior, es necesario tener al menos un año de datos para poder cuantificar el efecto de cada feriado. Por consiguiente, las tiendas de vestuario infantil no presentan valores en el regresor $Feriados_t$.

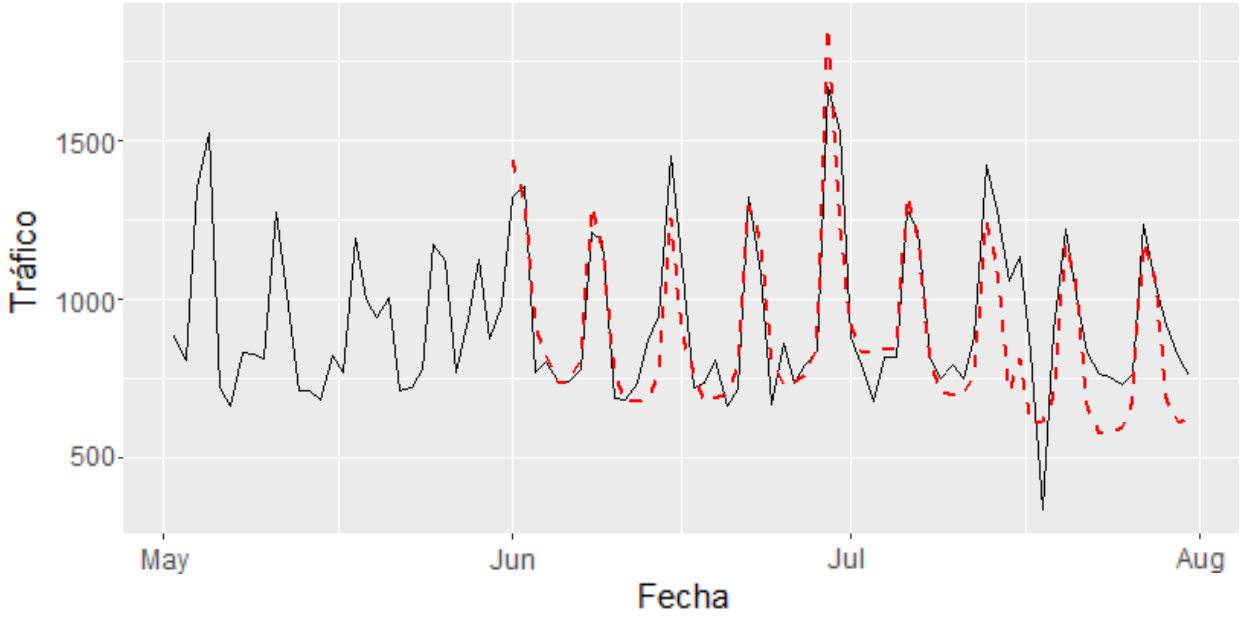
Posteriormente, si dicho festivo se encuentra en los días a predecir, el modelo incluye el efecto. Debido a que el modelo es una serie aditiva, el efecto también lo es. En conclusión, el modelo queda definido de la siguiente manera:

$$y_t = Trend_t + YearSeasonality_t + WeekSeasonality_t + Feriados_t + FirstDays_t + \varepsilon_t$$

Utilizando el modelo descrito, se realiza el pronóstico de las 54 tiendas. Para la tienda utilizada como ejemplificación, se obtiene el siguiente pronóstico (***Figura 20***).

Figura 20

Pronóstico Prophet II para una tienda de repuesto automovilístico. MAPE = 12.40%.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

A partir de la ***Figura 20***, se puede observar como el pronóstico cambia radicalmente el día festivo. En promedio, el MAPE correspondiente a las 54 tiendas es de 19.85%. Los resultados por tipo de tienda se presentan en la ***Tabla 21.***

Tabla 21

Resultados pronóstico Prophet II por tipo de tienda.

Tienda	MAPE	
	Dentro	Fuera
Vestimenta Infantil	**15.04%**	N/A
Repuestos Automóvil	10.70%	**12.78%**
Vestimenta Deportiva	16.54%	**27.21%**

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Es importante recalcar que los feriados se repiten anualmente y provocan diversos efectos en el tráfico. Por ejemplo, hay feriados que aumentan el tráfico en las tiendas mientras que hay otros que disminuyen el tráfico (normalmente los feriados irrenunciables).

Sin embargo, si bien no existe un efecto directo del regresor $Feriados_t$ en las tiendas de vestimenta infantil, el modelo logra diferenciar entre los días normales y festivos. Al poder discriminar entre dichos días, los componentes de tendencia y estacionalidad en la serie de tiempo no se ven afectados por el tráfico anómalo presenciado en los días festivos. No obstante, el modelo no puede replicar el efecto en los días festivos. Dado lo anterior, se puede determinar que, al poder diferenciar entre día normal y festivo, la precisión del pronóstico para las tiendas de vestuario infantil mejora en 3 puntos porcentuales con respecto al modelo que no diferencia por el tipo de día.

También, se puede observar que el efecto de los primeros cuatro días de cada mes corresponde a un aumento de tráfico en 53 de las 56 tiendas estudiadas. El aumento porcentual promedio es de 6.7%, con las tiendas de vestimenta infantil presentando la mayor variación con un aumento de 10.7% y las tiendas de repuestos presentando el menor aumento de 4.2%. Estos datos obtenidos tienen bastante sentido con la realidad, donde las tiendas de repuesto, al tener bienes urgentes, las personas no esperan recibir el sueldo para ir a comprar el repuesto del automóvil. Asimismo, en Chile las revisiones técnicas obligatorias deben ser realizadas en el mes que le corresponda a cada vehículo. Normalmente, la demanda por las revisiones aumenta a fin de mes por lo que, a su vez, aumenta el tráfico para los repuestos. Por otro lado, el pronunciado aumento del tráfico en las tiendas de vestuario infantil durante los primeros cuatro días del mes se puede deber a dos factores principales. Los productos de vestuario infantil no son urgentes y los padres suelen tener muchos gastos por tener un niño, lo cual los hace más susceptibles a esperar a comienzos de mes para realizar las compras respectivas.

7.1.8. Prophet III

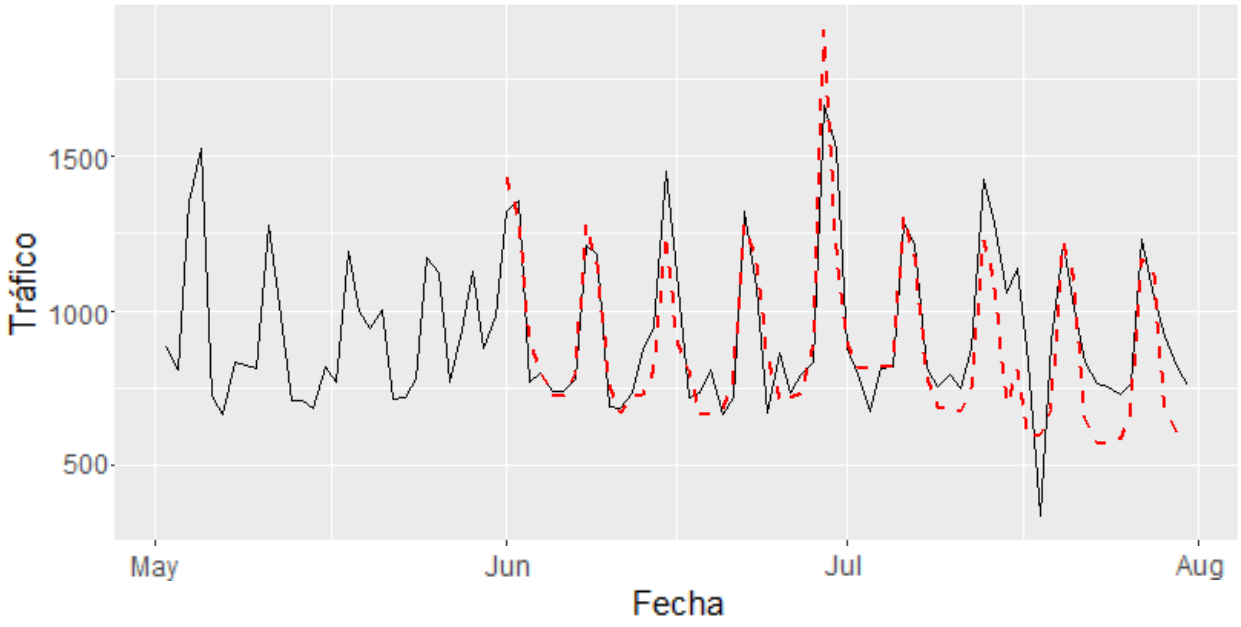
Según el estudio realizado por Martínez-de-Albéniz y Belkaid (2019), el comportamiento de los clientes se ve afectado por características climatológicas. En particular, se concluye que las precipitaciones afectan considerablemente el tráfico de los clientes. En particular, el efecto promedio para las 98 tiendas que fueron analizadas en dicho estudio es una disminución de 7.4% en las tiendas calle y un aumento de 5.2% en las tiendas *mall*. En consecuencia, se busca incluir en el modelo la precipitación diaria acumulada. Adicionalmente, se incluye como una variable binaria que indica si había presencia de precipitación para una determinada tienda y día en específico. En conclusión, el modelo queda definido de la siguiente manera:

$$y_t = Trend_t + YearSeasonality_t + WeekSeasonality_t + Feriados_t + FirstDays_t + Rain_t + \varepsilon_t$$

Para obtener los valores de $Rain_t$, se consulta los datos de las 816 estaciones pertenecientes a la Dirección Meteorológica de Chile y Dirección General de Aguas. Utilizando el modelo descrito, se realiza el pronóstico de las 54 tiendas. Para la tienda utilizada como ejemplificación, se obtiene el siguiente pronóstico (***Figura 21***).

Figura 21

Pronóstico Prophet III para una tienda de repuesto automovilístico. MAPE = 12.87%.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

En promedio, el MAPE correspondiente a las 54 tiendas es de 19.75% Los resultados por tipo de tienda se presentan en la ***Tabla 22***.

Tabla 22

Resultados pronóstico Prophet III por tipo de tienda.

Tienda	MAPE	
	Dentro	Fuera
Vestimenta Infantil	**14.92%**	N/A
Repuestos Automóvil	10.60%	**12.75%**
Vestimenta Deportiva	16.27%	**26.98%**

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Para que el modelo pueda considerar el efecto de la precipitación, se debe entregar el regresor $Rain_t$ con los valores en los conjuntos de entrenamiento y validación. Lo anterior significa que, para generar predicciones, es necesario predecir los valores diarios de $Rain_t$. Si bien la tecnología actual nos permite predecir con seguridad el clima, no es posible predecir más allá de 7 días. En consecuencia, al igual que el modelo 1, no se pueden generar pronósticos a largo plazo. Esto significa que no es útil para realizar decisiones estratégicas como por ejemplo la cantidad de productos a ordenar para la siguiente temporada, pero si para realizar cambios de dotación para la siguiente semana.

Además, es prescindible mencionar que la cantidad de estaciones es limitada, por lo que hay ciertas locaciones del territorio chileno que no poseen medidor de precipitación como es el caso de Coronel, que tiene 96.000 habitantes.

7.1.9. Árboles aleatorios

Con el aumento de la inteligencia artificial (IA), actualmente existen varias investigaciones que han utilizado modelos de aprendizaje supervisado para la predecir valores futuros. *Random Forest* es un algoritmo de aprendizaje supervisado, el cual combina un conjunto de modelos débiles para generar un modelo robusto.

Debido a que el algoritmo se basa en árboles de decisión, es mejor agrupar los días especiales en tres categorías: Irrenunciables, no irrenunciables y días especiales. Los días especiales corresponden a días no festivos, pero con un aumento considerable en las ventas como por ejemplo es el día de la madre. Al agrupar estos días, el algoritmo podrá detectar de mejor manera la diferencia de impacto en el tráfico para cada tipo.

Además, la precisión del algoritmo *random forest* depende en el ajuste de los parámetros y las variables seleccionadas, también conocidas como selección de características. Para la predicción de una serie de tiempo, es necesario determinar qué *lags* deben estar incluidos en el modelo. Si bien la inclusión de una mayor cantidad de variables aumenta la precisión del pronóstico, es importante mencionar que el tiempo de procesamiento del algoritmo aumenta considerablemente. Asimismo, incrementar el número k de *lags*, inevitablemente disminuye el conjunto de entrenamiento en k días. Por lo tanto, es necesario decidir una cantidad óptima k de *lags* a incluir en el modelo.

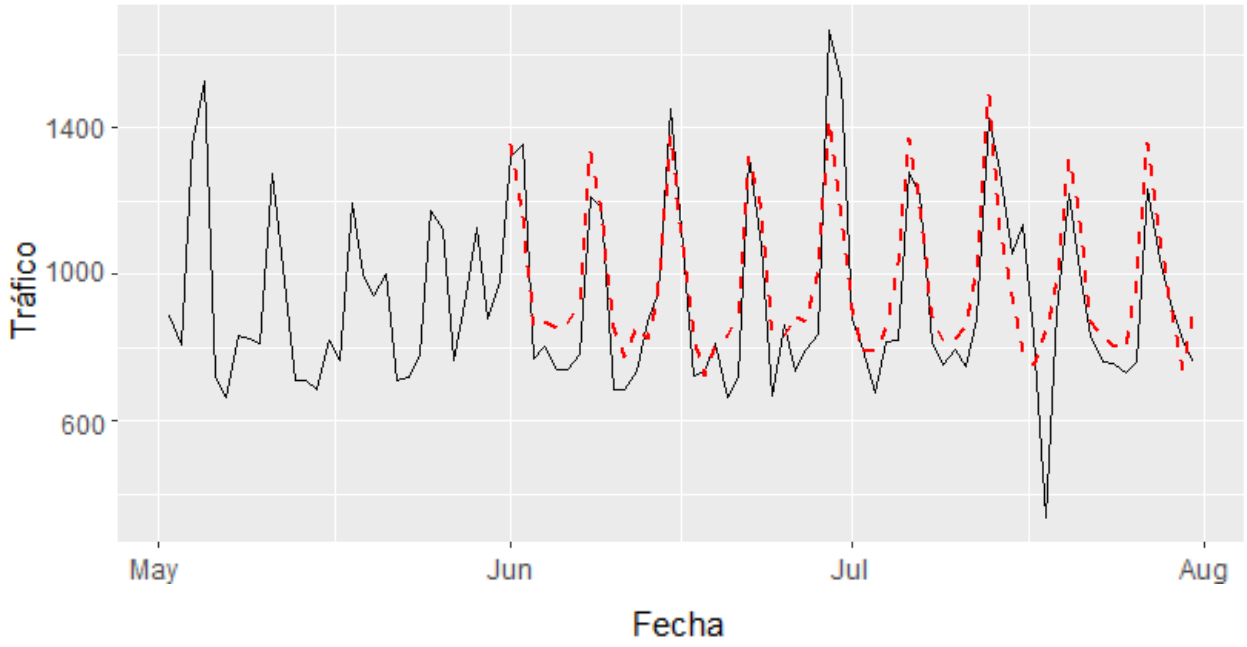
Se establece dos semanas de *lag* con el motivo de incluir correctamente la estacionalidad semanal. Si bien, la estacionalidad anual no es captada por el modelo, en la sección anterior fue comprobado que, en promedio, la estacionalidad semanal influye un 140% más que la estacionalidad anual.

No obstante, cabe señalar que el pronóstico es realizado en un computador con baja capacidad de procesamiento, por lo que los tiempos de procesamiento debiesen ser menores al descrito anteriormente.

A partir de las características mencionadas anteriormente, se generan las predicciones para las 54 tiendas utilizando 500 árboles aleatorios. Para la tienda utilizada como ejemplificación, se obtiene el siguiente pronostico (***Figura 22***).

Figura 22

Pronóstico Árboles aleatorios para una tienda de repuesto automovilístico. MAPE = 13.60%.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

En promedio, el MAPE correspondiente a las 54 tiendas es de 21.72%. Los resultados por tipo de tienda se presentan en la ***Tabla 23***.

Tabla 23

Resultados pronóstico árboles aleatorios por tipo de tienda.

Tienda	MAPE	
	Dentro	Fuera
Vestimenta Infantil	**18.64%**	N/A
Repuestos Automóvil	5.77%	**13.18%**
Vestimenta Deportiva	10.90%	**29.98%**

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

A partir de la ***Figura 22***, se puede observar que el modelo no incorpora completamente el efecto de las fechas especiales.

Como ya se ha mencionado, *random forest* tiene varias ventajas, más asimismo posee desventajas que se presentan en el desarrollo del modelo y son necesarias señalar. En primer lugar, es importante destacar que el algoritmo es una herramienta de modelado predictivo y no descriptiva. Esto significa que el modelo actúa como una caja negra, que no permite entender las relaciones en los datos. En segundo lugar, el tiempo de procesamiento es considerable, tardando 55 minutos en generar la predicción para las 54 tiendas.

7.2. Análisis comparativo de modelos

Luego de haber establecido los nueve modelos en la sección anterior, es necesario realizar un análisis comparativo entre ellos, con la finalidad de escoger el modelo que mejor se ajuste a los datos. La ***Tabla 24*** establece la comparación de rendimiento de los nueve modelos descritos y desglosado por tipo de tienda.

Tabla 24

Tabla comparativa de los nueve modelos predictivos.

Modelo	Vestuario Infantil		Repuesto Automóvil		Vestuario Deportivo	
	Dentro	Fuera	Dentro	Fuera	Dentro	Fuera
Naive I	N/A	30.37%	N/A	17.29%	N/A	30.55%
Naive II	N/A	N/A	N/A	19.85%	N/A	31.58%
Arima I	36.12%	N/A	20.17%	16.47%	40.79%	33.55%
Arima II	28.24%	N/A	20.06%	20.54%	31.49%	35.24%
Arima III	22.96%	N/A	17.56%	17.23%	23.71%	33.21%
Prophet I	16.81%	N/A	10.77%	12.79%	17.58%	27.85%
Prophet II	15.04%	N/A	10.70%	12.78%	16.54%	27.21%
Prophet III	14.92%	N/A	10.60%	12.75%	16.27%	26.98%
FR	18.64%	N/A	5.77%	13.18%	10.90%	29.98%

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

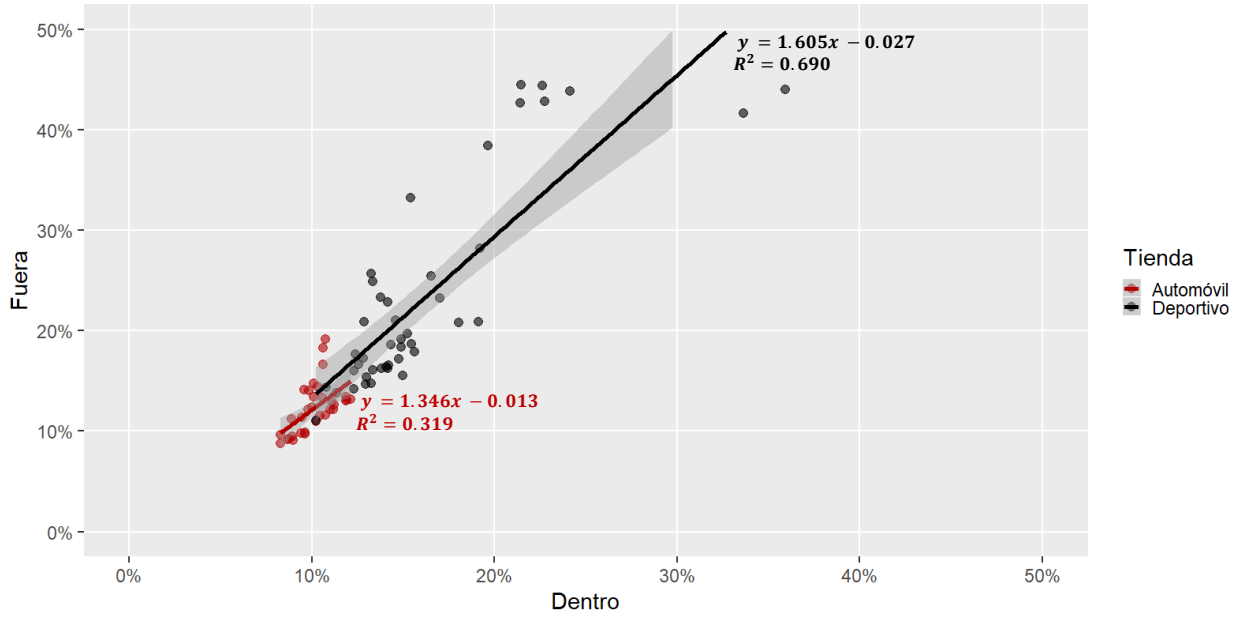
A partir de la figura anterior, se pueden destacar varios análisis. En primer lugar, se puede observar que a medida que aumenta la complejidad del modelo, en promedio se tiende a obtener predicciones más precisas. Lo anterior no se cumple completamente, pero se puede notar que los modelos ARIMA, Prophet y *random forest* son bastante mejor a los modelos *Naive* I y II, lo cual soporta la hipótesis establecida al comienzo de la sección. En consecuencia, es posible concluir que gran parte de la industria del *retail* no está utilizando las herramientas disponibles actualmente para poder entender de mejor manera al consumidor, con el fin de adecuar la dotación del personal, como también determinar la cantidad de stock óptima en inventario.

En segundo lugar, se puede observar que el modelo Prophet III es el que mejor predice fuera de la muestra en ambas tiendas (automóvil y deportivo). Prophet III reduce el MAPE del modelo de Olivares y Yung (2018) en 7.79 y 8.26 puntos porcentuales (en adelante, pp.) para las tiendas de automóvil y deportivo respectivamente. Lo anterior se puede deber a la inclusión de más variables relevantes (lluvia) y la incorporación de otras estacionalidades (mensual y anual). Asimismo, el modelo presenta una mejora en 13.32 pp. en el pronóstico de las tiendas de vestuario infantil. Por otro lado, el modelo árboles aleatorios presenta un excelente pronóstico dentro de la muestra para las tiendas de automóvil y deportivo, presentando un MAPE de 5.77% y 10.90% respectivamente. Debido a que el modelo escogido debe ser el que mejor se ajuste a los datos fuera de la muestra, se escoge Prophet III como el modelo con mayor precisión. Dicho modelo se utiliza en la próxima sección para cuantificar – en términos de dinero – la mejora del pronóstico en comparación con el modelo de Olivares y Yung (2018).

En tercer lugar, es interesante poder analizar la precisión de la predicción desglosado por tipo de tienda. Si bien podemos declarar que el modelo Prophet III pronostica con mayor precisión las tiendas de automóvil que las de vestimenta deportiva, se hace difícil incorporar las tiendas de vestimenta infantil debido a que no poseen pronósticos fuera de la muestra. En consecuencia, es necesario analizar el impacto entre los tipos de pronóstico, con la finalidad de extrapolar los datos de infantil para obtener un valor estimado de la precisión fuera de la muestra. Para lograr lo anterior, se grafica el MAPE dentro versus fuera de la muestra para todas las tiendas de automóvil y deportivo de los modelos Prophet (***Figura 23***).

Figura 23

Gráfico dentro vs fuera de la muestra para los modelos Prophet.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

A partir de la figura anterior, se puede observar una relación lineal entre el MAPE del pronóstico dentro de la muestra versus fuera de ésta, desglosado por el tipo de tienda (automóvil y deportivo). Además, las correlaciones de los datos con las rectas son de 0.319 y 0.690 para las tiendas de automóvil y deportivo respectivamente.

Para obtener un aproximado de la precisión para las tiendas de vestuario infantil, se extrapolan las dos relaciones lineales y se reemplaza $x = 14.92\%$. Debido a que se tienen dos rectas, se puede obtener un intervalo de valores posibles:

$$y = 1.346x - 0.013 \qquad y = 1.605x - 0.027$$

$$y = 23.91\% \qquad y = 27.36\%$$

Se obtiene que utilizando el modelo Prophet III, el pronóstico fuera de la muestra para las tiendas de vestuario infantil tendría una precisión entre 23.91% y 27.36%.

Finalmente, se propone desarrollar un modelo híbrido entre Prophet III y árboles aleatorios que pueda mejorar la precisión de los modelos por sí solos.

8. Impacto económico

Para poder mostrar el valor del presente trabajo, es necesario determinar el impacto económico al mejorar la predicción de tráfico. Si bien una buena predicción de tráfico posee varias ventajas económicas, las cuales fueron mencionadas en la sección 1.3, en este informe solo se analiza el impacto económico debido a la dotación eficiente de trabajadores. El análisis realizado se basa completamente en el trabajo realizado por Olivares y Yung (2018). Es importante mencionar que todos los modelos utilizados para la obtención de la utilidad esperada son de propiedad de Olivares y Yung. El único propósito en la utilización de estos algoritmos es para determinar el impacto económico de una mejor predicción en el tráfico de clientes, por lo que la explicación de los modelos utilizados a continuación no se encuentra en el alcance del presente informe. Dichos académicos utilizaron los mismos datos de las tiendas de vestuario infantil.

El objetivo de esta sección es obtener un valor económico – asociado a la dotación de trabajadores – del nuevo modelo de predicción llevado a cabo por el autor. Lo anterior se puede llevar a cabo calculando la diferencia de la utilidad esperada del modelo Prophet III con el modelo ARIMA I, generado por Lam et al. (1998).

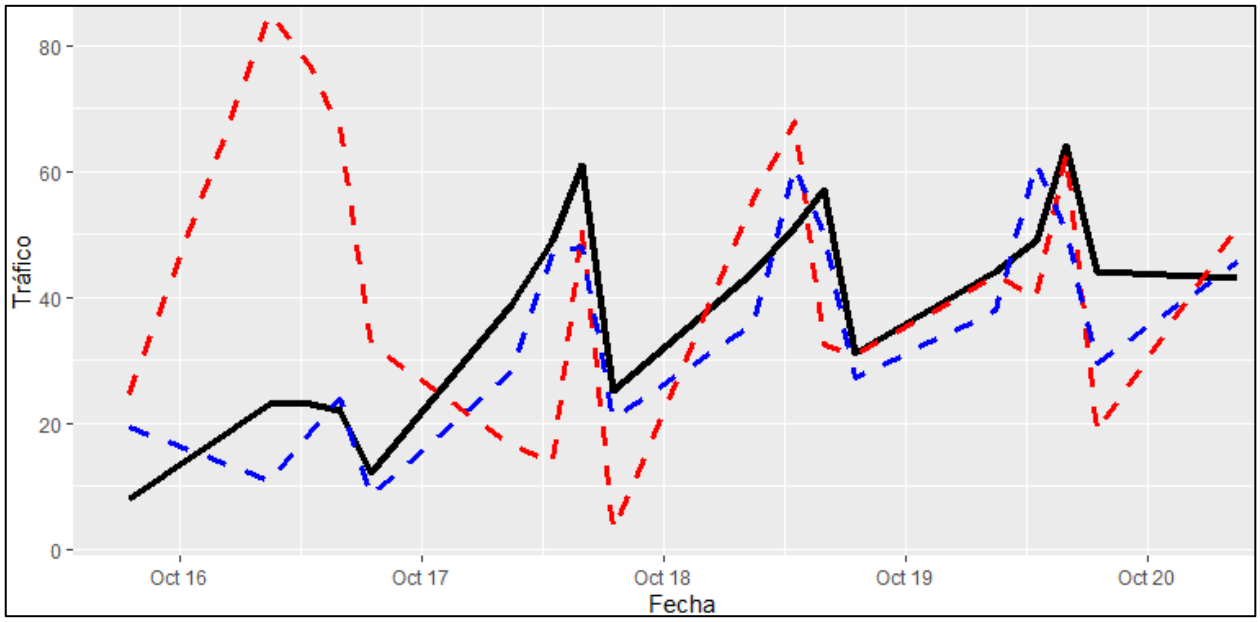
Sin embargo, es necesario mencionar que las predicciones realizadas por Olivares y Yung son por bloques de 3 horas (9am - 12pm, 12 - 3pm, 3 - 6pm y 6 – 9pm) y no por día como se presenta a lo largo del presente informe. Si bien las predicciones por día son útiles para decisiones operativas y tácticas, es obligatorio realizar predicciones en un intervalo de tiempo menor, ya que existe la posibilidad de incluir los trabajadores *part-time* que pueden trabajar media jornada. En consecuencia, se replica el modelo formulado por Lam et al. (1998).

Bajo este contexto, se predicen las 13 tiendas de vestimenta infantil para el mes de octubre 2016. Se obtiene un MAPE promedio de 55.21%. Asimismo, se realiza nuevamente el modelo Prophet III agregado en bloques de 3 horas para las tiendas de vestuario infantil, obteniéndose un MAPE de 25.70%, es decir, 29.51 puntos porcentuales menos que el modelo anterior.

La ***Figura 24*** muestra una comparación en el pronóstico de ambos modelos para la primera tienda durante el 16/10/2016 y 20/10/2016.

Figura 24

Ejemplo ARIMA I (rojo) y Prophet III (azul) para una tienda de vestimenta infantil.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Además, la ***Tabla 25*** muestra la comparación del MAPE para ambos modelos con las distintas ventanas de tiempo, 3 y 24 horas.

Tabla 25

Precisión de los pronósticos ARIMA I y Prophet III.

Modelo	3 horas	1 dia
Arima II	54.89%	28.24%
Prophet III	25.70%	14.92%

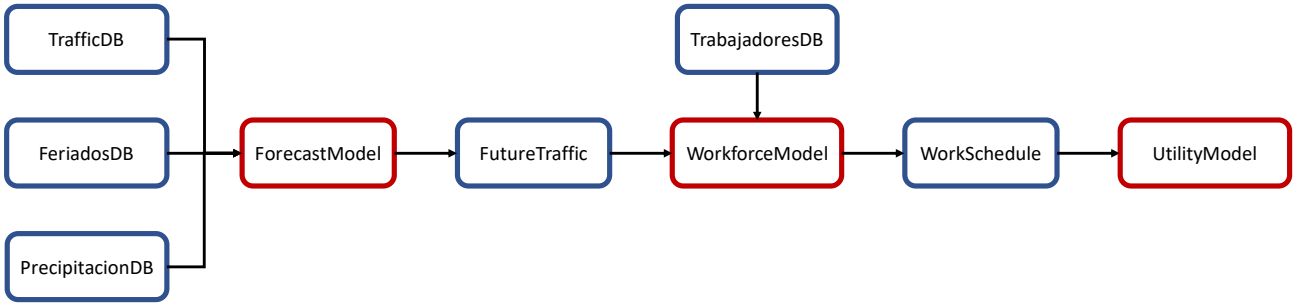
Fuente: Creación propia. Datos proporcionados por Intelligenxia.

A Partir de la ***Tabla 25*** se puede observar que la predicción diaria tiene un mayor MAPE para ambos modelos, aún cuando los modelos no varían. Al agregar los datos, la varianza y el ruido disminuye con respecto al nivel de la serie de tiempo. En consecuencia, se obtienen predicciones con mayor precisión y, por ende, menor MAPE.

Además, a partir de la ***Tabla 25*** se puede observar que el modelo Prophet III ajusta mejor que el modelo ARIMA II. Se plantea como hipótesis inicial que, a mayor poder predictivo, más precisa será la dotación de trabajadores. Esto a su vez implica mayores utilidades al largo plazo al minimizar la sobre y sub dotación de trabajadores. Para comprobar o refutar la hipótesis mencionada, se sigue el esquema de la ***Figura 25***, con el fin de obtener las utilidades esperadas.

Figura 25

Esquema de datos (azul) y modelos (rojo).



Fuente: Creación propia.

Para calcular el impacto económico, se siguen los siguientes pasos:

1. Se calcula ajusta función de ventas que depende del tráfico y cantidad de trabajadores, obteniéndose los betas de la regresión.

2. Se utiliza el pronóstico de tráfico obtenido en la sección anterior (FutureTraffic) para evaluar la función de ventas para cada bloque, día y cantidad de trabajadores.

3. Se calcula la combinación óptima de trabajadores y los tipos de turno (especificados en TrabajadoresDB), maximizando la utilidad (considerando el costo de trabajadores).

4. Utilizando nuevamente la función de ventas, se calcula la utilidad para el tráfico real y la dotación de trabajadores.

5. Finalmente, se calcula la diferencia en utilidad entre los pronósticos de Lam et al. (1998) y Olivares y Yung (2018).

En primer lugar, se calcula la función de ventas, correspondiente a la cantidad de transacciones (N_{it}), que se modela como un proceso de Poisson, utilizando la herramienta computacional STATA.

$$E(N_{it}) = T_{it} exp(\delta_i - \rho \log(T_{it}) + f(T_{it}, E_{it}; \beta) + Controls)$$

donde el término exponencial captura la conversión de tráfico en ventas. Dicha conversión incluye la saturación de tráfico ρ y el efecto de moderación del tráfico de clientes con el número de empleados en la tienda. El modelo se estima calculando la función de máxima verosimilitud. La ***Tabla 1 del Anexo 1*** muestra los parámetros estimados del proceso de Poisson.

En segundo lugar, se genera una tabla con las utilidades para cada bloque de 6 horas (2 por día) durante una semana de una tienda en particular. Además, se consideran todas las combinaciones de número de trabajadores, con un máximo de 7. La ***Tabla 2 del Anexo 1*** muestra un ejemplo, donde las columnas y las filas corresponden a los bloques (14) y la cantidad de trabajadores (7) respectivamente.

En tercer lugar, se establece la cantidad óptima de trabajadores para el tráfico pronosticado, el cual considera las utilidades calculadas anteriormente y el costo de los trabajadores. Para poder utilizar el modelo de dotación, es necesario definir los tipos de horarios y sus respectivas limitaciones. Se utilizan los mismos horarios especificados en el trabajo de Olivares y Yung (2018):

- A Tiempo Completo (FT): 5 días a la semana en jornada completa. En Chile, los trabajadores a tiempo completo no pueden trabajar dos domingos seguidos. El costo por semana es de $75.000 cada uno.

- Part Time (PT30): Normalmente equivale a 6 jornadas a la semana a medio turno. Sin embargo, también pueden ser tres jornadas a turno completo o una combinación entre turno medio y completo. El costo por semana es de $45.000 cada uno.

- Part Time (PT10): Normalmente equivale a trabajar una jornada completa a la semana, pero también pueden ser dos turnos con mitad de jornada. El costo por semana es de $16.667 cada uno.

En cuarto lugar, se calcula la utilidad de la dotación óptima calculada para un pronóstico de tráfico determinado, utilizando la función de ventas, en conjunto con el tráfico real. Además, se calculan los costos de los trabajadores para la tienda y semana analizada, con el fin de obtener la utilidad neta.

Finalmente, se repiten los pasos anteriores para los modelos Lam et al. (1998) y Olivares y Yung (2017), con el fin de obtener el valor económico del modelo predictivo desarrollado en el presente trabajo.

Para determinar el impacto económico, se utilizan 10 tiendas de vestuario infantil y la primera semana de octubre 2016, correspondiente a las fechas entre el 2016-10-02 y 2016-10-08, inclusive. La ***Tabla 26*** sintetiza los principales resultados obtenidos.

Tabla 26

Utilidades pronosticadas por tienda y modelo predictivo.

tienda	Utilidad ARMA	Utilidad Prophet	Delta	Delta Porcentual
1	$10,009,903	$10,251,228	$241,325	2.41%
3	$5,629,455	$5,642,475	$13,019	0.23%
4	$4,713,512	$4,806,437	$92,925	1.97%
5	$5,190,762	$5,247,683	$56,921	1.10%
7	$4,847,926	$4,903,794	$55,869	1.15%
8	$4,439,533	$4,469,057	$29,524	0.67%
9	$5,804,214	$5,889,797	$85,584	1.47%
10	$9,919,329	$10,217,439	$298,111	3.01%
12	$8,930,988	$9,024,987	$94,000	1.05%
13	$5,615,992	$5,593,777	-$22,216	-0.40%
Total	$65,101,614	$66,046,675	$945,062	1.45%

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

A partir de la ***Tabla 26***, se puede confirmar la hipótesis planteada anteriormente, ya que el modelo con mayor poder predictivo también obtiene la mayor cantidad de utilidad neta (Prophet). Se obtiene que, en promedio, el modelo predictivo desarrollado en la sección 7.1.8 (Prophet III) mejora las rentabilidades de las tiendas en 1.45%, al minimizar la sobre y sub dotación de personal.

9. Identificación de datos errados

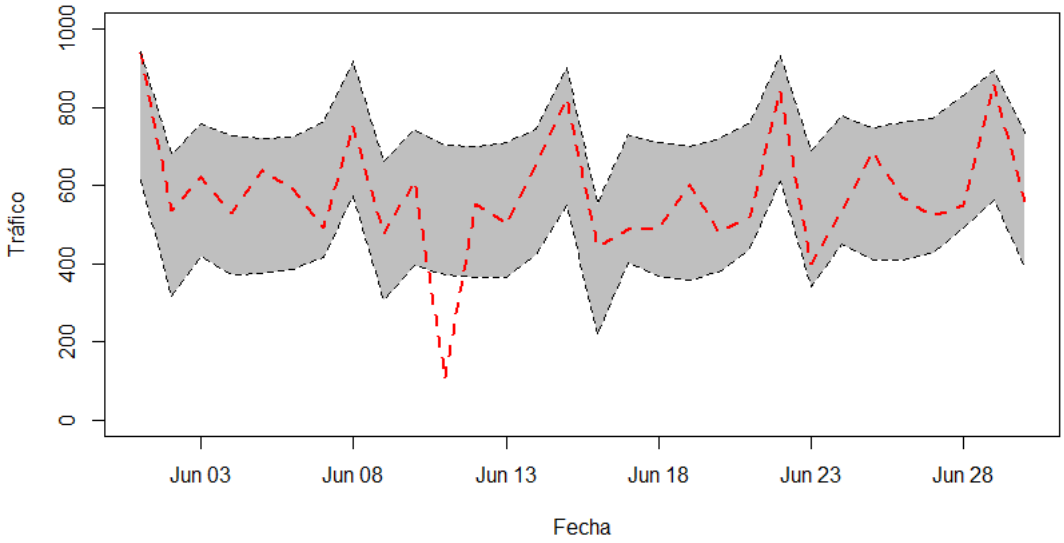
En la sección 6, se identificaron las tiendas con *data* faltante (tráfico nulo o menor a 5 el cual considera los posibles clientes fantasmas). Sin embargo, también es posible que una tienda registre una cantidad errada de tráfico para una determinada fecha. Lo anterior puede ocurrir debido a la falla del sensor, la caída del Wi-Fi en una tienda (por lo que los datos del sensor no son transmitidos a la base de datos) o un error en la recepción de los datos debido a una falla en el servidor donde se encuentra la base de datos. Dadas las problemáticas anteriores, es necesario tener un sistema automático que pueda identificar dichos errores para que puedan ser solucionados en el menor tiempo posible.

Para identificar la falla de un sensor, se utiliza los conocimientos del algoritmo de control estadístico de procesos. Lo anterior consiste utilizar técnicas estadísticas, lo que permite emplear criterios objetivos con el fin de distinguir datos anómalos significativos. Asimismo, al algoritmo mencionado anteriormente, se le agrega una variable de decisión que depende de las tiendas similares, utilizando las correlaciones entre éstas y la tienda que se estudia. El concepto preliminar es que si una tienda presencia un tráfico anómalo, y dicho comportamiento no ocurre en tiendas similares, se puede argumentar que existe una falla en la medición.

En mayor detalle, el concepto anterior se puede separar en dos objetivos de igual importancia. En primer lugar, se debe determinar para cada tienda, cuándo se presenta un dato anómalo. Para esto, se utiliza el modelo Prophet III, con el fin de obtener un intervalo de confianza al 95%. Los valores que se encuentren fuera del intervalo de confianza serán catalogados como anómalos (***Figura 26***).

Figura 26

Ejemplo visual para la identificación de datos errados.



Fuente: Creación propia. Datos proporcionados por Intelligenxia.

En segundo lugar, se debe determinar cómo comparar el dato anómalo con las tiendas similares, con el fin de decidir si la efectivamente el sensor está fallando o si bien, el efecto corresponde a un efecto común presente en otras tiendas. Las tiendas similares son aquellas que poseen una correlación fuerte entre ellas. Utilizando la ***Tabla 5***, se considera correlación fuerte cuando el coeficiente es mayor a 0.7.

Como ya fue mencionado, un dato anómalo se define como un valor fuera del intervalo de confianza pronosticado el cual, valga la redundancia, puede ser mayor o menor. Para comprobar que efectivamente se trata de una falla en el sensor, las tiendas similares a la tienda en estudio no pueden poseer un dato anómalo con la misma tendencia (mayor o menor) para el mismo día a ésta.

Para una mayor comprensión del algoritmo, se precisan algunas variables. Sea i una tienda, se define y_{it} como el valor real de la tienda i en el periodo t. Asimismo, se definen los intervalos de confianza del pronóstico para una tienda i en el periodo t: L_{it} y U_{it}, equivalentes al límite inferior y superior respectivamente. Se define la variable binaria x_{it} que identifica cuando existe un dato anómalo en la tienda i para el período t.

$$x_{it} \begin{cases} 1 & y_{it} < L_{it} \quad \vee \quad U_{it} < y_{it} \\ 0 & L_{it} \leq y_{it} \leq U_{it} \end{cases}$$

Además, sea j otra tienda perteneciente al conjunto de tiendas K_i similares a i, entonces el fallo de un sensor se puede confirmar a partir de una anomalía identificada en la tienda i para el periodo t $(x_{it} = 1)$, a través de la siguiente expresión:

- Si $\boldsymbol{y_{it} > U_{it}}$: $\quad \forall \boldsymbol{j} \in \boldsymbol{K_i}, \quad \boldsymbol{y_{jt} \leq U_{jt}}$
- Si $\boldsymbol{y_{it} < L_{it}}$: $\quad \forall \boldsymbol{j} \in \boldsymbol{K_i}, \quad \boldsymbol{y_{jt} \geq L_{jt}}$

Bajo esta línea, para poder determinar cuándo un sensor falla, es necesario que la tienda i posea un conjunto K_i no vacío de tiendas similares. De las 54 tiendas, 50 poseen al menos una tienda similar. De las cuatro tiendas sin dicho conjunto, tres corresponden al rubro de vestimenta infantil, mientras que una corresponde al rubro de repuestos automovilísticos. Para efectos de la investigación, solamente se aplica el modelo para estas 50 tiendas mencionadas anteriormente.

Para demostrar el funcionamiento, se utiliza el algoritmo descrito anteriormente por un mes en todas las 50 tiendas. Los meses son los mismos que fueron pronosticados. En caso de las tiendas de vestuario deportivo y de repuestos, se considera solamente el primer mes, es decir, junio 2018 y 2019 respectivamente. En total, se tienen 1.500 observaciones que el algoritmo debe clasificar en dos categorías: dato normal (N) o anómalo (A). A partir de las 1.500 observaciones, el algoritmo identifica 38 fallas. 10 para las tiendas de repuesto automovilístico, 1 y 28 para las tiendas de vestuario infantil y deportivo respectivamente.

Al correr el algoritmo en la herramienta R, se obtiene una tabla informando todas las fallas, especificando el tipo de tienda, numero de tienda, fecha y si corresponde a un valor sobre o bajo el intervalo de confianza (**Tabla *27***).

Tabla 27

Informe de potenciales fallos en los sensores.

empresa	tienda	fecha	Comentario
Repuesto Automóvil	1	2019−06−30	Subida considerable del tráfico
Repuesto Automóvil	3	2019−06−22	Baja considerable del tráfico
Repuesto Automóvil	3	2019−06−24	Baja considerable del tráfico
Repuesto Automóvil	3	2019−06−27	Baja considerable del tráfico
Repuesto Automóvil	5	2019−06−01	Baja considerable del tráfico
Repuesto Automóvil	6	2019−06−07	Baja considerable del tráfico
Repuesto Automóvil	6	2019−06−11	Baja considerable del tráfico
Repuesto Automóvil	7	2019−06−05	Baja considerable del tráfico
Repuesto Automóvil	11	2019−06−14	Subida considerable del tráfico
Repuesto Automóvil	17	2019−06−13	Baja considerable del tráfico
Vestuario Infantil	1	2016−10−09	Subida considerable del tráfico
Vestuario Deportivo	4	2018−06−30	Baja considerable del tráfico
Vestuario Deportivo	13	2018−06−17	Baja considerable del tráfico
Vestuario Deportivo	18	2018−06−01	Subida considerable del tráfico

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Al momento de identificar la falla en el periodo t del sensor perteneciente a la tienda i, se realiza la misma imputación de datos utilizado en la sección 6.3. Lo anterior, mejora el pronóstico, ya que se elimina un dato incorrecto. Esto implica que el modelo de pronóstico no considere datos errados que puedan "ensuciar" el pronóstico.

Es necesario destacar la importancia de un buen modelo predictivo, ya que éste afecta directamente en la calidad del proceso de control estadístico que identifica las fallas en los sensores. Además, un algoritmo eficiente es capaz de señalar cuando se presenta una anomalía, pero al mismo tiempo, debe tener la capacidad de no señalar cuando no existe anomalía alguna.

El actual algoritmo posee tres limitaciones importantes a destacar. En primer lugar, si la tienda i tiene un conjunto vació K_i, no es posible determinar si efectivamente el sensor está fallando o si se debe a una variable externa como un corte de energía en la ciudad o manifestaciones masivas que obligar a cerrar la tienda. En segundo lugar, debido a que las tiendas se encuentran en distintas ciudades del país, es posible que un sensor de una tienda i tenga mediciones anómalas pero que se deba a variables externas encontradas solamente en esa ciudad. Por ejemplo, un corte de energía en la tienda i, pero no en la tienda j. En tercer lugar, el modelo no funciona en los casos que el límite inferior L_{it} sea negativo. Esto ocurre en tiendas con un bajo promedio de tráfico, pero con alta volatilidad. Esto provoca que el intervalo de confianza sea bastante amplio, obteniéndose un límite inferior negativo. Para las tiendas de vestuario infantil y deportivo, se presenta un L_{it} negativo en 22.1% y 21.7% de los casos. Por otro lado, las tiendas de repuesto no poseen L_{it} negativos.

Además, el proceso de control estadístico no logra identificar la falla en el caso de que varios sensores presenten el mismo problema. Por ejemplo, un corte de energía en una ciudad con varias tiendas, el cual resulta en valores menores al límite inferior permitido (siempre y cuando este límite no sea negativo). Si bien el algoritmo no identificara la falla, si serán informados como datos anómalos. Asimismo, identificar una falla general como la descrita en el párrafo anterior es bastante fácil debido a su importancia y tamaño. Es muy difícil que, a simple vista, la falla pase desapercibida.

Bajo la misma línea, es importante destacar que el algoritmo sirve solo para un conjunto K_i no vacío. En caso de que la tienda i no tenga tiendas similares, se puede considerar que el sensor está fallando cuando se detecte anomalía (sin tener que corroborar con las tiendas similares).

10. Conclusiones

Al momento de predecir series temporales, es de suma importancia determinar las variables exógenas que generan variabilidad. En el caso de los clientes de tiendas físicas, se comprueba que el comportamiento varía dependiendo del día del mes, clima y de los días feriados. En primer lugar, estudios de Pagel (2016) y Macleod (2014) comprueban que los consumidores suelen visitar más las tiendas a comienzos de cada mes, ya que corresponden a los primeros días posteriores de recibir el sueldo. En segundo lugar, Martínez-de-Albéniz y Belkaid (2019), comprobaron que el comportamiento de los clientes se ve afectado por características climatológicas, obteniendo un aumento de 5.2% en las tiendas *mall* y una disminucion de 7.4% en las tiendas calle. En tercer lugar, existen una serie de días especiales que no siguen el patrón predictivo del modelo. El efecto del día especial depende de su naturaleza. En este trabajo los feriados son de dos tipos, renunciables e irrenunciables. Los primeros implican, en promedio, un aumento del tráfico de 19%, mientras que el segundo implica que la tienda se encuentre cerrada, es decir, con tráfico nulo.

Prophet es una herramienta predictiva capaz de producir pronósticos de alta calidad para datos de series temporales que tienen estacionalidades múltiples con crecimiento lineal o no lineal. Además, Prophet identifica automáticamente quiebres de tendencia. Dicha herramienta es de fácil uso y extremamente útil para personas que recién están iniciando a pronosticar series temporales. El modelo Prophet I, el más simple formulado en el presente trabajo, obtiene una mayor precisión que el resto de los modelos formulados (*naive*, ARIMA y *random forest*).

Los sensores que captan el tráfico no poseen un 100% de precisión, por lo que es necesario identificar los días en que presentan fallas, ya que, dichas mediciones no son íntegras. Para identificar dichas fallas, se utiliza un control estadístico de procesos, utilizando como límites de aceptación el intervalo de confianza de las predicciones realizadas con el modelo Prophet III. Además, se genera un algoritmo de imputación de datos efectiva para los días en que el sensor falla, utilizando como referencia tiendas similares a través de la correlación de *Pearson*. Se obtiene un MAPE de 15.24%.

El modelo Prophet III, que incluye días festivos y datos de precipitación obtiene una precisión de 25.70% versus 54.89% del modelo formulado por Lam et al. (1998), en las predicciones por hora. Dicha diferencia se puede traducir en una asignación de personal optimizada, minimizando la sobre dotación y sub dotación. Se obtiene que el modelo predictivo Prophet III, en promedio, incrementa la rentabilidad en 1.45%.

La presente investigación presenta una serie de limitaciones. En primer lugar, los pronósticos calculados son realizados dentro de la muestra, lo cual inevitablemente sobre ajusta el modelo predictivo. En segundo lugar, el estudio se limita solamente a tiendas de *retail* en Chile e industrias de vestuario infantil, vestuario deportivo y repuestos de automóvil.

Finalmente, la presente investigación tiene varias aristas en las cuales se puede mejorar. En primer lugar, es posible que existan otras variables exógenas que puedan afectar la serie temporal. Por ejemplo, indicadores de la economía local. En segundo lugar, es posible cuantificar la precisión del control estadístico de procesos, al realizar simulaciones de fallas hipotéticas. En tercer lugar, es posible obtener predicciones de tráfico más robustas con modelos híbridos. Se propone realizar un modelo ponderado entre Prophet III y *random forest*.

11. Bibliografía

Brown, R. (1959). *Statistical forecasting for inventory control*, New York: McGraw-Hill.

Box, G. y Jenkins, M. (1970). *Time Series Analysis, Forecasting, and Control*. San Francisco: Holden-Day, Inc.

Breiman, L. (2001). *"Random Forests"*. *Machine Learning*. 45 (1): 5–32.

Carreno, J., y Madinaveitia, J. (1990). *A modification of time series forecasting methods for handling announced price increases.* International Journal of Forecasting, 6, 479–484.

Chuang, H., Oliva, R. y Perdikaki, O. (2016). *Traffic-based labor planning in retail stores.* Production and Operations Management, 25(1) 96–113.

Columbia Business School. *On Payday, Consumers Feel a License to Spend.* [En línea] [consulta: 12 de febrero de 2020].
<https://www8.gsb.columbia.edu/articles/ideas-work/payday-consumers-feel-license-spend>

Cottrell, M., Girard, B., Girard, Y., Mangeas, M. y Muller, C. (1995). *Neural modeling for time series: a statistical stepwise method for weight elimination*. IEEE Transactions on Neural Networks 6 (6), 1355–1364.

Gardner, E. (1985). *Exponential smoothing: the state of the art*. Journal of Forecasting, 4, 1–38.

Gardner, E. (1993). *Forecasting the failure of component parts in computer systems: A case study*. International Journal of Forecasting, 9, 245–253.

Grubb, H. y Masa, A. (2001). *Long lead-time forecasting of UK air passengers by Holt-Winters methods with damped trend.* International Journal of Forecasting, 17, 71–82.

Harvey, A. y Peters, S. (1990). *Estimation procedures for structural time series models.* Journal of Forecasting 9, 89–108.

Ho, T (1995). *Random Decision Forests.* Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

Holt, C. (1957). *Forecasting seasonals and trends by exponentially weighted averages.* International Journal of Forecasting, 20, 5–13.

Hu, M. (1964). *Application of the adaline system to weather forecasting.* Master Thesis, Technical Report 6775-1, Stanford. Electronic Laboratories, Stanford, CA, June.

Intelligenxia. *Nosotros.* [En línea] [consulta: 9 de octubre de 2019] <http://intelligenxiabg.com/nosotros/>

Intelligenxia. *Cómo Funciona.* [En línea] [consulta: 9 de octubre de 2019] <http://intelligenxiabg.com/como-funciona/>

Khoshgoftaar, T., Golawala, M., y Hulse, J. (2007). *An empirical study of learning from imbalanced data using random forest*. IEEE Computer Society, 2, 310–317.

Kolmogorov, A. (1941). *Stationary sequences in Hilbert space* (in Russian). Moscow University Mathematics Bulletin, No. 6.

Lam, S., Vandenbosch, M. y Pearce, M. (1998). *Retail sales force scheduling based on store traffic forecasting.* Journal of Retailing, 74(1) 61–88.

Lever, G. (2019). *Tendencias del comercio electrónico en Chile*. Centro Economía Digital CCS.

LinkedIn. *Artificial Neural Networks Advantages and Disadvantages*. [En línea] [consulta: 23 de abril de 2020] <https://www.linkedin.com/pulse/artificial-neural-networks-advantages-disadvantages-maad-m-mijwel/>

López, J. (2018). *Análisis de Series de Tiempo. Pronóstico de demanda de uso de aeropuertos en Argentina al 2022*. Universidad Tecnológica de Buenos Aires.

Mani, V., Kesavan, S. y Jayashankar, M. (2015). *Estimating the impact of understaffing on sales and profitability in retail stores.* Production and Operations Management, 24(2) 201–218.

Miller, T. y Liberatore, M. (1993). *Seasonal exponential smoothing with damped trends. An application for production planning*. International Journal of Forecasting, 9, 509–515.

Olivares, M. y Yung, D. (2018). *Labor planning and shift scheduling in retail stores using customer traffic data.* 1–31.

Quenouille, M. (1957). *The Analysis of Multiple Time-Series*. London.

R Development. *Prophet Forecasting at Scale*. [En línea] [consulta: 10 de octubre de 2019] <https://research.fb.com/prophet-forecasting-at-scale/>

Ríos, G. (2008). *Series de Tiempo.* Santiago: Universidad de Chile. Recuperado de: <https://www.ucursos.cl/ingenieria/2010/1/CC52A/1/material_docente/bajar?id_material=296003>

Rosas, A. y Guerrero, V. (1994). *Restricted forecasts using exponential smoothing techniques.* International Journal of Forecasting, 10, 515–527.

Rumelhart, D., Hinton, G. y Williams, R. (1986). *Learning representations by backpropagating errors.* Nature 323 (6188), 533–536.

Segal, M. (2004). *Machine Learning Benchmarks and Random Forest Regression.* UCSF: Center for Bioinformatics y Molecular Biostatistics. Recuperado de: <https://escholarship.org/uc/item/35x3v9t4>

Sharda, R. y Patil, R. (1992). *Connectionist approach to time series prediction: An empirical test*. Journal of Intelligent Manufacturing 3, 317–323.

Shumway, R. y Stoffer, D. (2006). *Time Series Analysis and Its Applications. With R Examples.* Davis, California. 2da Edición. Springer. Recuperado de: <http://db.ucsd.edu/static/TimeSeries.pdf>

Tang, Z., Almeida, C. y Fishwick, P. (1991). *Time series forecasting using neural networks vs Box-Jenkins methodology*. Simulation 57 (5), 303–310.

Tang, Z. y Fishwick, P. (1993). *Feedforward neural nets as models for time series forecasting*. ORSA Journal on Computing 5 (4), 374–385.

The Drum. *Infographic: The habits of British spending around payday.* [En línea] [consulta: 11 de febrero de 2020]. . <https://www.thedrum.com/stuff/2014/08/27/infographic-habits-british-spending-around-payday>

The Professionals Point. *Advantages and Disadvantages of Random Forest Algorithm in Machine Learning.* [En línea] [consulta: 25 de abril de 2020]. <http://theprofessionalspoint.blogspot.com/2019/02/advantages-and-disadvantages-of-random.html>

Thomas M. (1983). *Short Term Forecasting, An introduction to the Box-Jenkins Approach*. John Wiley y Sons.

Weigend, A., Huberman, B. y Rumelhart, D. (1990). *Predicting the future: A connectionist approach*. International Journal of Neural Systems 1, 193–209.

Williams, D. y Miller, D. (1999). *Level-adjusted exponential smoothing for modeling planned discontinuities*. International Journal of Forecasting, 15, 273–289.

Winters, P. (1960). *Forecasting sales by exponentially weighted moving averages*. Management Science, 6, 324–342.

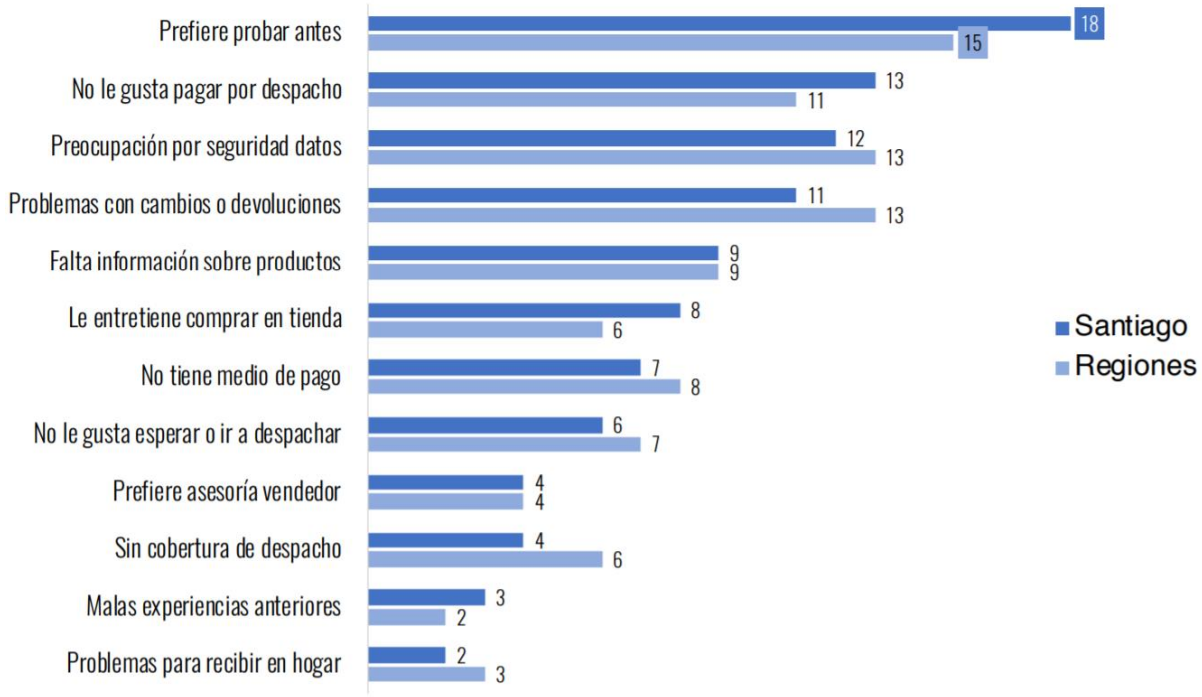
Yenidogan, I., Cayir, A. y Kozan, O. (2018). *Bitcoin Forecasting Using ARIMA and Prophet*. Conference on Computer Science and Engineering. Recuperado de: <https://www.researchgate.net/publication/329400738_Bitcoin_Forecasting_Using_ARIMA_and_PROPHET/citations>

12. Anexos

12.1. Anexo I: Figuras

Figura 1

Barreras a la compra online en Chile.



Fuente: Centro Economía Digital CCS: *"Tendencias del comercio electrónico en Chile"*.

12.2. Anexo II: Tablas

Tabla 1

Estimación de los parámetros del modelo de ventas (Poisson).

Variable	β	se
constante	-1.3820	0.0827
ln(tráfico)	0.9950	0.0109
ratio	-0.0482	0.0078
ratio2	0.0026	0.0006
tienda 1	0	0
tienda 3	-0.5540	0.0139
tienda 4	-0.5640	0.0130
tienda 5	-0.7460	0.0134
tienda 8	-0.6650	0.0140
tienda 9	-0.4840	0.0144
tienda 10	-0.6200	0.0129
tienda 11	-0.5580	0.0152
tienda 13	-0.5980	0.0131
tienda 14	-0.6660	0.0140
tienda 15	-0.3050	0.0145
dayofweek 0	0	0
dayofweek 1	-0.1720	0.0246
dayofweek 2	-0.1360	0.0244
dayofweek 3	-0.1010	0.0242
dayofweek 4	-0.0524	0.0240
dayofweek 5	0.0245	0.0236
dayofweek 6	0.1580	0.0113
block 1	0	0
block 2	0.4070	0.0180
block 3	0.3760	0.0185
block 4	0.3970	0.0184
weekend 1 block 1	0	0
weekend 1 block 2	-0.1190	0.0244
weekend 1 block 3	-0.1440	0.0239
weekend 1 block 4	-0.2060	0.0236
lluvia	0.0020	0.0115
feriado	0.0571	0.0536

$N = 12{,}240, \quad R^2 = 0.644, \quad AIC = 65402, \quad BIC = 66286$

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

Tabla 2

Utilidades por bloque de 6 horas y día para distintas cantidades de trabajadores (máximo 7).

Día	Domingo		Lunes		Martes		Miércoles		Jueves		Viernes		Sábado	
	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM
1	1,181,254	1,104,275	426,283	450,088	462,585	463,027	489,495	443,410	499,796	500,690	594,033	707,885	1,626,412	1,573,640
2	1,262,836	1,166,176	447,952	474,072	486,727	487,993	514,269	464,677	524,911	526,199	626,546	750,476	1,731,073	1,636,447
3	1,309,873	1,211,334	456,425	483,775	496,196	498,129	523,850	473,024	534,657	536,381	639,476	768,735	1,799,511	1,701,893
4	1,337,664	1,239,234	460,913	488,974	501,218	503,568	528,907	477,452	539,806	541,813	646,365	778,708	1,840,983	1,744,254
5	1,355,804	1,257,815	463,690	492,210	504,325	506,955	532,028	480,192	542,986	545,185	650,639	784,973	1,868,367	1,773,029
6	1,368,534	1,271,007	465,576	494,416	506,437	509,264	534,146	482,055	545,145	547,481	653,547	789,270	1,887,713	1,793,687
7	1,377,947	1,280,836	466,941	496,016	507,966	510,940	535,677	483,403	546,706	549,145	655,655	792,400	1,902,081	1,809,192

Fuente: Creación propia. Datos proporcionados por Intelligenxia.

12.3. Anexo III: Descripción Base de datos

La base de datos posee los siguientes campos:

- Tienda: Corresponde al nombre de la tienda.
- Empresa: Tipo de tienda. Repuesto automóvil, vestuario infantil y vestuario deportivo.
- Tipoevento: Tipo de evento que detecta el sensor. Entrada o salida.
- Fecha: Fecha del evento
- Hora: Hora del evento
- Tráfico: Cantidad de personas contabilizadas