Instituto Tecnológico
y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Maestría en Ciencia de Datos

Pronóstico de costos por obligaciones de garantías a corto plazo

---

TRABAJO RECEPCIONAL que para obtener el **GRADO** de
MAESTRO EN CIENCIA DE DATOS

Presenta: **DANIEL FRANCISCO NUÑO ÁLVAREZ**

Director: **DR. JAIME EMMANUEL ALCALÁ TEMORES**

Tlaquepaque, Jalisco, Junio de 2023.

AGRADECIMIENTOS

Dedicado a mi familia y amigos cercanos, en especial a mi madre, mejor amigo y mentor, quienes me han apoyado a ser mejor. Su amor y aliento inquebrantables han sido mi luz guía a lo largo de mi viaje académico. Siempre estaré agradecido por sus sacrificios y su confianza en mí.

RESUMEN

Cada mes la organización de finanzas de HP Inc. debe proveer un estimado de los costos mensuales de operación para atender obligaciones de garantía a corto plazo, con un máximo de 6 meses hacia el futuro. Actualmente, el proceso es intensivo en tiempo y en labor, dejando mucho que desear a la precisión.

El enfoque de este trabajo es mejorar la precisión, usando datos de HP Inc., por medio de métodos de aprendizaje de máquina. Se evalúan modelos como series de tiempo, máquinas de vectores de soporte, y redes neuronales. Al final se determina cual modelo se ajusta mejor con respecto a ciertas métricas y se discuten brevemente los resultados para 1 de las 14,725 series de tiempo en el alcance del proyecto.

La automatización del proceso para entrenar, evaluar y producción de pronósticos es también parte del reporte y del trabajo de implementación ya que es una actividad que se debe realizar mensualmente con los nuevos datos. La división organizacional requiere una división geográfica, por línea de costo, y por línea producto.

TABLA DE CONTENIDO

1. INTRODUCCIÓN

Resumen: *Cada mes la organización de finanzas de HP Inc. debe proveer un estimado de los costos mensuales de operación para atender obligaciones de garantía a corto plazo, con un máximo de 6 meses hacia el futuro. Actualmente el proceso es intensivo en tiempo y en labor, dejando mucho que desear a la precisión.*

1.1. CONTEXTO

El pronóstico de costos asociados al atendimiento, soporte y reparación de productos por obligaciones de garantías es multidisciplinario y complejo porque incluye:

- Análisis financiero: desde el punto de vista de finanzas corporativas que involucra entendimiento de estado de resultados, balance general, costos variables y costos fijos. Si el problema lo requiriere tener entendimiento de la estructura organizacional, impuestos o inventarios, por ejemplo, para hacer frente a cambios en la dinámica como afectarían los costos.
- Investigación de operaciones: desde el punto de vista de manufactura que involucra saber cuántos productos de la compañía existen en el periodo que cubre la garantía y la probabilidad de falla de dichos productos en el futuro para luego asignarle un costo.
- Ciencia de datos: desde el punto de vista de métodos de inteligencia artificial y aprendizaje de máquina que involucra el estudio, en este caso, el análisis de datos históricos como series de tiempo y el ajuste de un modelo capaz de pronosticar el futuro.

En finanzas corporativas existen dos metodologías para pronosticar estados de resultados o flujo de efectivo:

- Estratégico: basado en objetivos con enfoque mediano y largo plazo (el año actual a más de tres años). Fuertemente dependiente en el conocimiento humano. Por ejemplo, introducción de nuevos productos o costos no vistos antes.
- Táctico: basado en datos de tus estados financieros e indicadores externos. Enfoque en corto plazo (hasta 12 meses adelante). Yves R. Sagaert, El-Houssaine Aghezzaf, Nikolaos Kourentzes, Bram Desmet lo estudian particularmente para el pronóstico de ventas utilizando modelos estadísticos y variables macroeconómicos para mejor la precisión.[1]

En el trabajo de Wu y Akbarov [2] describen detenidamente desde una perspectiva de manufactura como pronosticar los reclamos de garantías a través de máquinas vector de soporte.

La competencia M5 es una aplicación de pronóstico de ventas minoristas con el objetivo de producir los pronósticos puntuales más precisos para 42,840 series de tiempo que representan las ventas unitarias jerárquicas de la empresa minorista más grande del mundo,

así como para proporcionar las estimaciones más precisas de la incertidumbre de estas previsiones. Por lo tanto, la competencia consistió en dos desafíos paralelos: pronóstico de Precisión e Incertidumbre.

Makridakis, Spiliotis, y Assimakopoulos [3] describen los antecedentes, la organización y las implementaciones de la competencia, y presenta los datos utilizados y sus características. Sirve como ejemplo de la industria para desafíos que requieren gran cantidad series de tiempo.

A. David Lainder y Russell D. Wolfinger [4] ofrecen orientación para abordar estas dificultades y proporcionar un marco que maximice las posibilidades de predicciones que se generalicen bien. Las técnicas que, usadas para la validación cruzada, el aumento y el ajuste de parámetros se han utilizado para ganar varios concursos importantes de pronóstico de series de tiempo, incluido el concurso M5 Forecasting Uncertainty y la serie Kaggle COVID19 Forecasting.

Danny Yuan, de UBER, explica intuitivamente la transformación rápida de Fourier y la red neuronal recurrente (seq2seq). Explora cómo estos conceptos juegan un papel crítico en el pronóstico de series de tiempo cíclicas.[5]

1.2. JUSTIFICACIÓN

En la esfera empresarial, surge una imperante exigencia de acortar sustancialmente el tiempo asociado al proceso decisional, sea este ejecutado de manera plenamente automatizada o semiautomatizada, con el fin último de lograr una mejoría ostensible en la calidad de las decisiones adoptadas, en aras de alcanzar una precisión superior. Ante este panorama, se torna imprescindible la implementación de enfoques basados en el aprendizaje automático, con especial énfasis en el pronóstico de series temporales, como una vía sumamente prometedora y altamente pertinente para abordar esta problemática y alcanzar resultados sustanciales.

1.3. PROBLEMA

Estimar los costos de garantías mensualmente es un problema para el cual aún no existe una solución automatizada. Estas garantías son de las computadoras e impresoras de uso comercial y personal vendidos de HP Inc. en todo el mundo. Geográficamente comprende 3 regiones y 8 mercados:

- América
  - Norte América
  - América Latina
- Europa, África y Medio Oriente
  - Europa Central

  - Europa Sur
  - Europa Noreste
- Asia Pacífico
  - China
  - Asia Mayor
  - India

En tipo de producto comprende 4 grandes segmentos:

- Computadoras
  - Comercial (empresariales)
  - Consumo (uso personal)
- Impresoras
  - Comercial (empresariales)
  - Consumo (uso personal)

El objetivo es crear una solución que pueda proveer una precisión, al menos, igual a las soluciones actuales, pero sin la bruma, el trabajo y el tiempo que conlleva hacerlo mes con mes. Idealmente será completamente automática, supervisada, online, pero hay consideraciones que no están capturadas en los datos, como información de partes altamente defectuosas, problemas en la cadena de suministro o inversiones.

Estas estimaciones en conjunto de otra información o estimaciones proporcionadas por otras organizaciones tienen tres propósitos principales que se usan internamente:

- Estimación del flujo de efectivo.
- Estimación de los estados financieros de la empresa.
- Responsabilidad a los altos ejecutivos

Parte de la visión de HP Inc. es la innovación digital e internamente transformar la forma en que trabajamos. El métrico principal es la precisión de la predicción evaluado mes con mes, es decir la diferencia entre predicción y real. El punto de referencia es la precisión de la solución actual. Adicionalmente, métricas relevantes son (1) cuántos días de laborales se puede reducir para la entrega de la predicción. Si ahora tarda un ciclo de 10 días en entregar entonces que tarde menos de 10 días. Y (2) cuántas horas de trabajo se reducen mes con mes, trabajo en horas por trabajador para entregar la predicción de gastos y costos.

Actualmente esta tarea tiene un costo inerte a la labor de todos los que participan que teóricamente puede reducirse con una nueva implementación. La solución no debe canjear precisión por costo, sino que, por lo menos, la precisión debe ser la misma.

Los costos y gastos se reportan mes con mes y se componen de costos regionales, gastos globales, y reservas y amortizaciones. Los costos globales son en su mayoría fijos relacionados a empleados o inversiones. Las reservas y amortizaciones responden a ahorros hechos para cubrir los costos basados en las ventas. Los costos regionales corresponden a costos fijos de

empleados, pero también a gastos variables operativos como partes de repuesto, cadena de suministro, logística, trabajo de ingenieros en la reparación, y llamadas de asistencia.

La siguiente lista corresponde a las grandes categorías de los tipos de costos:

- Total Warranty Expense
  - Region Owned Expense
    - Variable Expense
      - Contact Center
      - Delivery
      - Supply Chain
      - Other Repair Cost
    - Repair OH Expense
      - Contact Center OH
      - Delivery OH
      - Supply Chain OH
    - Other Warranty Expense
  - Worldwide Owned and Allocated Expense
    - CS HQ Owned and Allocated
      - CS HQ
      - CS Investments
    - GBU Owned and Allocated
      - GBU Owned and Allocated
  - Net Reserve Expense
    - Accrual for Shipments
    - Amortization

Ver sección 5.1 para una mayor explicación.

Para gastos de variables de contact center necesitamos saber tres cosas:

- $V = cantidad\ de\ unidades\ vendidas\ en\ un\ periodo.$
- $L = porcentaje\ de\ unidades\ vendidas\ con\ fallas.$
- $V * L = cantidad\ esperada\ de\ llamadas.$
- $C_{llamada} = costo\ promedio\ por\ llamda.$

$$CC_{variable} = V * L * C_{llamada}$$

Para gastos variables de reparación necesitamos saber tres cosas:

- $V = cantidad\ de\ unidades\ vendidas\ en\ un\ periodo.$
- $R = porcentaje\ de\ unidades\ vendidas\ que\ necesiten\ reparación.$
- $V * R = cantidad\ esperada\ de\ reparaciones.$
- $C_{reparacion} = costo\ promedio\ de\ reparación.$

$$repair_{variable} = V * R * C_{reparacion}$$

Para los costos fijos (overhead) necesitamos saber dos cosas:

- $E = cantidad\ de\ empleados.$
- $C_{empleado} = costo\ promedio\ por\ empleado.$

$$overhead = E * C_{empleado}$$

Estas ecuaciones son relevantes cuando consideramos un modelo de regresión para explicar y pronosticar las series de tiempo. Modelos multivariables que incluyen datos de ventas monetarias, unidades vendidas, intervenciones, llamadas, conteo de personal y otras posibles variables operativas.

Otro de los requisitos es la granularidad en la geografía (mercado) y el tipo de producto, lo que agrega complejidad al proceso por que los costos de un segmento y mercado terminan siendo diferentes. Son 8 mercados, 86 tipos de productos y 25 líneas de costo. La tabla 1 y figura 1 muestra la cantidad de series de tiempo requeridas por nivel.

Las suposiciones del problema son las siguientes:

- Tiene tendencia.
- Tiene estacionalidad.
- Es autorregresivo, los valores previos tienen un impacto en los siguientes valores.
- Un modelo de regresión tendría un mejor resultado que un modelo de series de tiempo univariada.
- Es un proceso estocástico porque hay costos no previstos o considerados deterministas.
- Los números reportados no son perfectos por errores humanos, cambios operativos, contables y de sistemas.
- Datos más recientes y entendimiento del modelo de negocio son más importantes para los pronósticos al futuro.
- Un modelo explicativo de cada línea de costos es más importante que los datos históricos. El pronóstico a futuro de variables operativas es vital para una buena precisión.

Tabla 1

Número de series por nivel de agregación			
Nivel id	**Nivel descripción**	**Nivel agregación**	**Número series**
1	Gasto total	Total	1
2	Gasto total para cada región	Región	3
3	Gasto total para cada mercado	Mercado	8
…	…	…	…
5	Gasto total para cada mercado y negocio	mercado – negocio	16
…	…	…	…
10	Gasto total para cada mercado y línea de prod.	mercado - producto	720
…	…	…	…
14	**Línea de costo i, para cada mercado y prod.**	**costo - mercado - producto**	**14,725**
Total			41,795

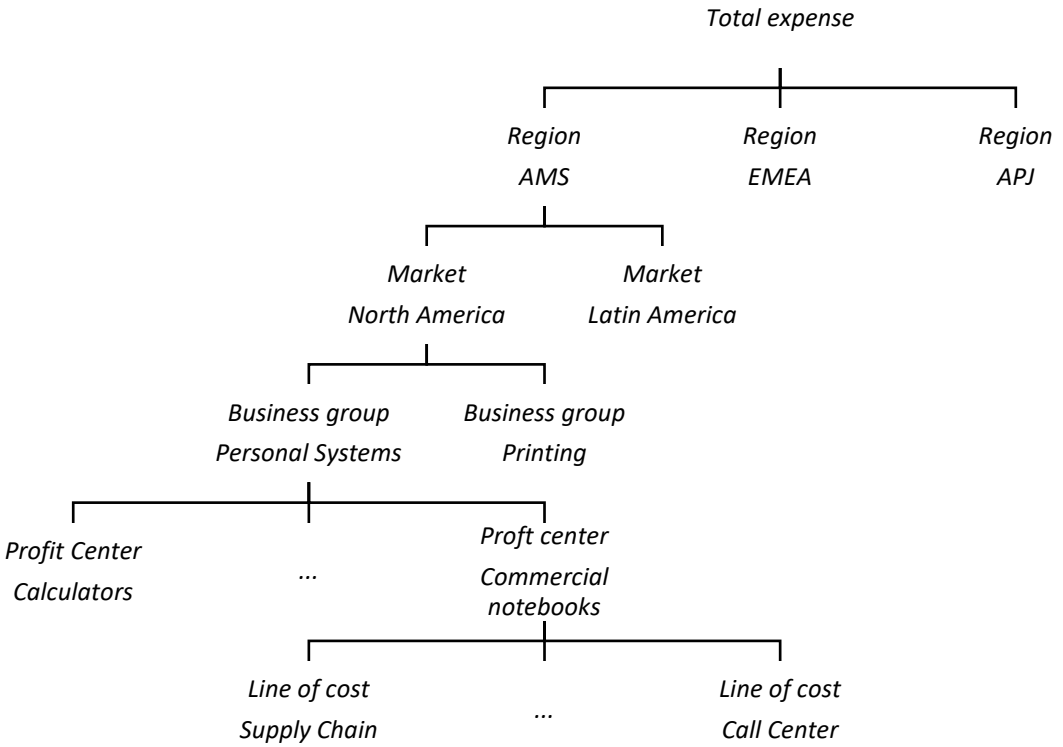


Figura 1. Agrupación y jerarquía de las series de tiempo que conforman el gasto total de garantías.

1.4. OBJETIVOS

1.4.1. OBJETIVO GENERAL:

Construir un modelo, o los modelos necesarios, para la organización de finanzas de HP Inc. que ayuden a pronosticar eficientemente a corto plazo los costos operativos.

1.4.2. Objetivos Específicos:

1. Evaluar y comparar diferentes modelos que mejoren la precisión, pero no sobreentrene en los datos históricos.
2. Utilizar el error relativo para medir la precisión.
3. Confirmar normalidad y homocedasticidad de los residuales.
4. Medir el tiempo de ejecución de cada algoritmo con los mismos recursos computacionales.
5. Usando los puntos 2, 3 y 4 decidir qué modelo es el mejor para cada serie.
6. Crear un proceso que permita analizar todas las series de tiempo de forma eficiente.
7. Proveer cada mes los pronósticos correspondientes.

2. METODOLOGÍA

Resumen: *El enfoque de este trabajo es mejorar la precisión por medio de métodos de aprendizaje máquina. Se evalúan modelos como series de tiempo, máquinas de vectores de soporte y redes neuronales.*

2.1. DESCRIPCIÓN DE LOS DATOS

Los datos para utilizar son los costos mensuales de cada una de las variables que componen operativamente la organización de garantías. Estos costos monetarios al ser divididos por mercado y por línea de productos estamos hablando de 14,725 series de tiempo. En cuanto al rango, los datos disponibles son de noviembre 2016 a noviembre 2022.

Los datos financieros son recolectados del Libro Contable. Datos operativos son recolectados de diferentes sistemas dependiendo de la región o el tipo de producto. Los datos son consolidados en una base de datos, por lo tanto, fuera del proceso siguiente descrito se codificó los valores para proteger la confidencialidad de HP Inc.

Estos datos son propiedad y confidenciales de HP Inc. y son usados por mi persona como empleado y bajo guía de mi jefe con la intención de mejorar el proceso.

La mayor parte del análisis, entrenamiento y resultados es basada en una serie de tiempo ya que el proceso sería el mismo para todas las demás. La serie analizada es:

- Línea de costo: Cadena de suministro.
- Mercado: Norte América.
- Producto: Computadoras personales.

Como variables exógenas tenemos a la disposición los siguientes datos internos:

- Ingresos: total de ventas en el periodo observado.
- Número unidades vendidas: total de unidades vendidas, computadoras, por ejemplo, en el periodo observado.
- Número de llamadas: Total de llamadas de asistencia en el periodo observado.
- Número de reparaciones: Total de reparaciones en el periodo observado.
- Base instalada anualizada: Base instalada es la cantidad de unidades, computadoras, por ejemplo, que aún se encuentran bajo una obligación de ser reparada en caso de falla.
- Tasa de intervenciones anualizada: cantidad de intervenciones sobre base instalada. Representa la cantidad de intervenciones por cada unidad. 0.2% significa que se reparan 2 computadoras por cada 1000.

El periodo de garantías es variable entre 3 y 12 meses dependiendo del tipo de producto. Por ejemplo, una computadora puede tener 12 meses de garantía a partir de la venta, por lo tanto, el ejercicio y costos asociados a la asistencia o reparación sucede después y en su mayoría cercana a la expiración de la garantía.

Las métricas operativas número de llamadas, reparaciones, base instalada y tasa de intervenciones anualizadas sirven para explicar los costos como fue descrito en la sección pasada.

Se utilizó el lenguaje de programación python y las librerías pandas[6], matplotlib[7], scipy[8], statsmodels[9], StatsForecast[10], scikit-learn[11], keras[12], tensorflow[13].

2.2. ANÁLISIS EXPLORATORIO

2.2.1. BENCHMARK DE PRECISIÓN

Primero vamos a analizar el problema y establecer un punto de referencia sobre la precisión, comparando los pronósticos con los resultados históricos. Los datos van de la siguiente manera:

Para cada mes existen n estimaciones de costos pasados, que pueden ser expresados como un vector:

$C =$ costo
$t =$ periodo
$Ca_t =$ costo subindice t, costo del periodo
$Cf_t =$ costo subindice t, costo del periodo
$n =$ número de periodos pasados

$$flash = \{Cf_{t-1}, Cf_{t-2}, Cf_{t-3}, \dots, Cf_{t-n}\}$$

El valor pronosticado es conocido como flash. El valor real, también llamado *actual.*

$$actual = Ca_t$$

El vector de error o desviación para cada periodo sea la diferencia del valor actual y cada uno de los valores del vector flash sobre el valor actual.

$$error = \{\frac{Ca_t}{Cf_{t-1}} - 1, \frac{Ca_t}{Cf_{t-2}} - 1, \dots, \frac{Ca_t}{Cf_{t-n}} - 1\}$$

De aquí podemos calcular el valor esperado y desviación estándar del error, lo cual determina nuestro punto de referencia.

De forma matricial, cada fila es un periodo de la forma que incluye el costo real y cada una de las estimaciones pasadas:

$$\{actual, flash\}$$

$$\{Ca_t, Cf_{t-1}, Cf_{t-2}, Cf_{t-3}, \dots, Cf_{t-n}\}$$

Definiendo $n = 6$ obtenemos la siguiente matriz.

Tabla 2

mes	t	t-1	t-2	t-3	t-4	t-5	t-6
3/1/2022	$12,354,729	$11,599,840	$11,599,840	$11,897,590	$12,395,410	$15,204,280	$15,204,280
4/1/2022	$9,569,062	$11,478,910	$11,478,910	$11,478,910	$11,821,550	$12,094,640	$15,237,640
5/1/2022	$9,526,362	$9,526,362	$10,917,230	$11,511,180	$11,902,590	$11,902,590	$11,902,590
6/1/2022	$10,159,004	$10,175,940	$10,175,940	$10,916,780	$11,499,050	$11,935,460	$11,935,460
7/1/2022	$6,882,127	$10,192,290	$10,192,290	$10,192,290	$10,986,240	$11,572,250	$12,000,350

La siguiente tabla evalúa el promedio de los errores:

Tabla 3

t-1	t-2	t-3	t-4	t-5	t-6
-3%	-3%	-6%	-10%	-16%	-21%

La siguiente tabla evalúa la desviación estándar de los errores.

Tabla 4

t-1	t-2	t-3	t-4	t-5	t-6
14%	16%	15%	17%	14%	13%

Lo que estamos viendo es que, en promedio, para el periodo inmediato posterior hay un error promedio de 3%. Que no es malo, pero podría ser mejor ya que, además de los problemas antes mencionados, el 3% significa 300 mil dólares (sin mencionar la serie aquí analizada representa solo 15% del total de todos los gastos operativos). El error es significativamente mayor en 6 periodos al futuro con una media de 21%.

La desviación estándar es constante en *t-1* hasta *t-6*, lo cual indica que los errores son constantes. Es constante que el error de *t-1* sea alrededor de 3% y que el error de *t-6* sea alrededor de 21% en la misma magnitud.

Es común la sobreestimación de los gastos, un término llamado *sandbagging*. En términos de estimación de flujo de efectivo y obligaciones, es igualmente malo la sobreestimación que la que subestimación. Discutiendo con expertos de finanzas, de operaciones y explorando los datos, se ha llegado a dos conclusiones que ya se asumían posibles.

En el transcurso de los 5 años de datos que se tienen, han ocurrido cambios operacionales que afectan la estructura de costos: la estrategia de atención a distancia, la forma de reparación, los tipos de productos, o el uso de proveedores terciarios. La estructura de costos sigue cambios operacionales que dificulta el análisis histórico, y más aún pronósticos, ya que existe poca información o comentarios que expliquen estos cambios del pasado ni que documenten los cambios futuros.

Existen discrepancias y atípicos en los datos debido a los cambios en los sistemas financieros, la separación de HP, cambios en las líneas de producto, agrupación de países que afectan las regiones, cambios en los procesos contables, rotación de analistas financieros y errores de los analistas financieros.

Los datos, la estructura contable, terminan siendo imperfectos y deficientes para modelos que dependan 100% de los datos, en este caso datos históricos. Por ejemplo, las siguientes observaciones de todos los tipos de costos:

- Los gastos Delivery OH y Supply Chain OH son nuevos. Empezaron a registrarse en noviembre de 2020 entonces solo hay 24 observaciones.
- Delivery subtipos de costos, llamados Direct e Indirect, empezaron a registrarse en noviembre de 2020 entonces solo hay 24 observaciones.
- Tipo de productos de consumer no registra Delivery porque por el tipo de reparación no es necesaria.
- CS HQ Owned and Allocated subtipo de gastos, llamados IT POA y Rapid and Radical, solo ocurren una vez por mes.
- Los costos de Contact Center por muchos años estuvieron rezagados por un mes. En *t* se registraba la actividad correspondiente, pero en *t+1* los costos. No estaban a la par pero a partir de 2022 ya están alineados.
- Los cartuchos de tinta y toner son un tipo de producto llamados supplies y existen para impresoras comerciales y consumo (uso personal). En nuestra organización no existe supplies comercial pero sí para consumo. Todos los costos y métricas relacionados a supplies comercial no tienen sentido y tienen que ser eliminados para reducir el ruido.
- Otros negocios tienen su propio soporte de garantías.
- Porque somos una compañía y existen economías de escala, en su mayoría todos los costos son centralizados, independientemente del tipo de producto o país, aunque si exista actividad como reparaciones, ventas o base instalada por país. Esto hace más difícil la estimación en los niveles más bajos

2.2.2. ANÁLISIS DESCRIPTIVO

Los datos disponibles son 75 periodos, desde noviembre 2016 hasta enero 2022. El periodo fiscal para HP comienza en noviembre y termina en octubre. Los trimestres comienzan en noviembre, febrero, mayo y agosto. Todos los valores son numéricos monetarios que representan el gasto o costos incurridos en el periodo.

En conjunto de todas las series de tiempo, es interesante analizar lo siguiente:

- Gráficas de serie de tiempo.
- Tendencia y estacionalidad a lo largo del año en los 12 meses.
- Clustering de series de tiempo.
- Evaluar la predictibilidad usando CV^2 y ADI.

Cada serie muestra sus propias características: algunas son suaves, mientras que otras son intermitentes y erráticas. Siguiendo las recomendaciones de Syntetos & Boylan [14] para cada serie se calcula CV^2 (coeficiente cuadrado de variación de la demanda cuando ocurre) que representa la irregularidad de la demanda, y ADI (intervalo promedio inter-demanda) nota la intermitencia en el tiempo.

$$ADI = \frac{número\ total\ de\ periodos}{número\ de\ periodos\ demandados} \qquad CV^2 = \frac{desviación\ estandar}{promedio}$$

En base a estas 2 dimensiones, la literatura clasifica los perfiles de demanda en 4 categorías diferentes:

- Demanda suave (ADI < 1,32 y CV^2 < 0,49). La demanda es muy regular en tiempo y en cantidad. Por lo tanto, es fácil de pronosticar y no tendrá problemas para alcanzar un nivel de error de pronóstico bajo.
- Demanda intermitente (ADI >= 1,32 y CV^2 < 0,49). El historial de demanda muestra muy poca variación en la cantidad demandada pero una gran variación en el intervalo entre dos demandas. Aunque los métodos de pronóstico específicos abordan demandas intermitentes, el margen de error de pronóstico es considerablemente mayor.
- Demanda errática (ADI < 1,32 y CV^2 >= 0,49). La demanda tiene ocurrencias regulares en el tiempo con altas variaciones de cantidad. La precisión de su pronóstico sigue siendo inestable.
- Demanda grumosa (ADI >= 1,32 y CV^2 >= 0,49). La demanda se caracteriza por una gran variación en cantidad y en tiempo. En realidad, es imposible producir un pronóstico confiable, sin importar qué herramientas de pronóstico utilice. Este tipo particular de patrón de demanda es impredecible.

Para análisis individual, mostrado en este reporte, es importante observar lo siguiente para cada serie de tiempo:

- Gráfica de serie de tiempo.
- Grafica distribución de valores.
- Valores atípicos y nulos.

- Sesgo.
- Homocedasticidad.
- Normalidad.

La cadena de suministros representa 47% de todos los gastos. Constituye a el costo más grande. Además, que es un costo variable, indirectamente dependiente de las ventas anteriores, que tan defectuoso sea el producto, lo costoso de la parte dañada, costos operativos de tener inventario, impuestos, y otros.

En la figura 2, la serie no muestra tendencia clara, se muestra un gran incremento de costos con respecto a meses previos en el último trimestre del 2019 y una gran caída coincidentemente con el declarado inicio de la pandemia, marzo y abril 2020, y por lo tanto la disminución de actividades. Los costos más altos son en 2021 que se puede relacionar con las garantías ejercidas de productos comprados entre 6 y 12 meses previos, (12 meses previos aumentó muchísimo la venta de computadoras). Además, que hubiera un exceso de inventario porque las cadenas de suministro estaban paralizadas y se tuvo que comprar en exceso para no interrumpir el servicio y se tuviera que vender o desechar posteriormente.

En la figura 2 y con las anotaciones de la figura 4 se aprecian tres cambios de tendencia. El primero entre el inicio y hasta octubre 2019. El segundo entre octubre 2019 y noviembre 2021. El tercero entre noviembre 2021 y Julio 2022. También en las mismas figuras se aprecian dos periodos de magnitudes de variación antes y después de febrero 2020.

En la figura 2, no se aprecia alguna estacionalidad por trimestre o año. Lo que es común es que un incremento comparado con el periodo anterior le procede un decremento; un cambio porcentual positivo es muy probable que le siga un cambio porcentual negativo. A pesar de no mostrar una clara estacionalidad, podría ser modelado razonablemente como una función continua y periódica expresada como la suma de otras series periódicas y lineales [15].

Hay ciertos conocimientos que pueden relacionarse a los datos observados en variables exógenas y otros que no (información de proveedores de servicios) y que la mejor solución sea una entrada manual de los expertos.

Para esta serie no existen valores faltantes, pero para las series que sí existan valores faltantes se consideran dos opciones:

1. La serie inicia después de enero 2017: En ese caso no se considera como un valor faltante.
2. La serie tiene faltantes después de haber iniciado: Después de comenzar la serie, muchos los valores atípicos son precedidos de un valor nulo (0), significa que en el valor nulo no existe registro de gastos y por lo tanto el siguiente mes es mucho más alto. Se podría decir que es lo correspondiente del mes pasado más el mes actual, un mes no hay gasto registrado y al siguiente es lo correspondiente a dos meses. Esa es la teoría, pero no se puede estar seguro de que siempre sea el caso.

El histograma en la figura 3 muestra sesgo positivo por lo que la transformación Box-Cox será útil. Los datos de esta serie, y muchas de las otras analizadas, no muestra patrón en tendencia ni estacionalidad. Es posible que el ajuste y pronósticos de la serie de tiempo sea poco útil por si sola. Modelos que consideren volatilidad no constante en el tiempo, autorregresivo y eventos especiales como indicadores binarios serán los candidatos.

Figura 2: Serie de tiempo

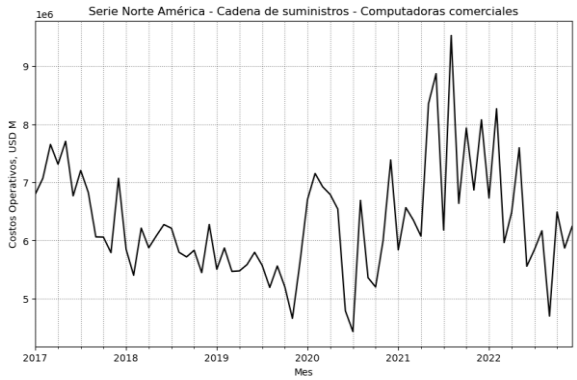


Figura 3: Histograma

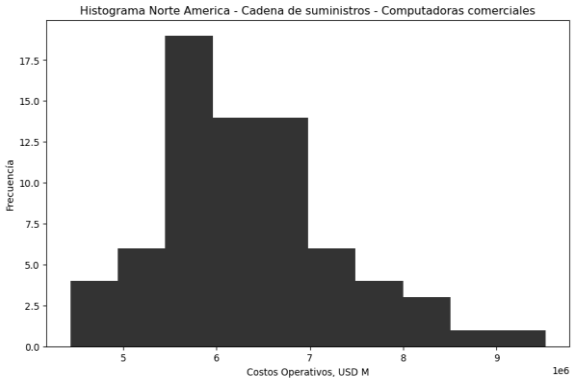


Figura 4: Serie de tiempo con media y desviaciones estándar

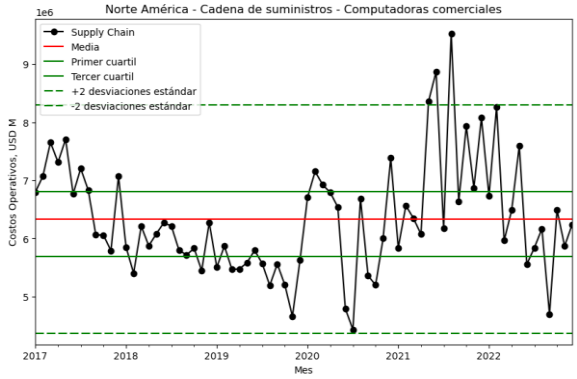
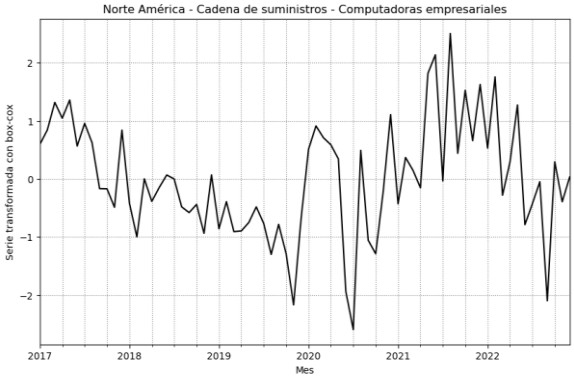


Figura 5: Serie transformada con Box-Cox



2.2.3. VALORES ATÍPICOS

Analizando una sola serie y con la intención de mantener los datos reales, la estrategia para cuantificar los datos atípicos en los modelos será usar las fechas mencionadas en el párrafo anterior como variables exógenas que no afecten el modelo generalizado.

Cuando sea el momento de analizar automáticamente todas las series de tiempo entonces se utilizará el proceso descrito por el Banco de Pagos Internacionales para estandarizar, agrupar y detectar valores atípicos [16].

Matemáticamente y con el soporte de la figura 4, se definiría un atípico como dos desviaciones estándar más menos la media del grupo seleccionado.

La estrategia para tratar esos valores atípicos es marcar la fecha como variable exógena que afecta el modelo (usando one-hot encoding siendo un binary indicator). Usualmente referenciados como *vacaciones* o *eventos especiales*. En el caso que existieran valores nulos se tratarían como eventos especiales.

2.2.4. TRATAMIENTO DEL SESGO

Para disminuir el sesgo y la varianza se utilizó la generalización Box-Cox ya que los valores son estrictamente positivos [17]. La figura 5 son los valores de la serie de tiempo después de aplicar la transformación Box-Cox.

Los valores resultantes de esta particular serie de tiempo son entre -3 y 3. El sesgo corrigió de 0.82 a -0.0035 usando un lamba de -0.71. Hasta aquí los datos que serán usados para entrenar están transformados y normalizados.

2.3. DESCRIPCIÓN DE LOS MODELOS

Uno de los cálculos que actualmente se hacen en la empresa para pronosticar es con promedios móviles simples. Con los modelos de series de tiempo, aunque no sean considerados modelos de aprendizaje automático buscamos algo que generalice, no sobreajuste, que sea más sofisticado y que obtenga mejores resultados que el benchmark.

El primer modelo es una generalización **SARIMAX** (seasonal autoregressive moving average with exogenous regressors) que mejor se ajusten a la serie [18]. Siendo una serie de tiempo, primero podemos analizar sin ninguna variable exógena. Siguiente, podemos agregar los periodos atípicos como variables para tratar de mejorar la precisión, y subsecuentemente, se usarán las variables mencionadas en la descripción de datos:

- Ingresos.
- Número unidades vendidas.
- Número de llamadas.
- Número de reparaciones.
- Base instalada anualizada.
- Tasa de intervenciones anualizada

Los siguientes modelos son un problema de regresión en el que se pueden usar las variables exógenas y, además, valores retrasados de la misma serie de tiempo.

El segundo modelo es **SVR** (support vector regression). Este modelo es interesante porque, por definición es un modelo regularizado que evita el sobreajuste [19].

El tercer modelo es una red neuronal **LSTM** (Long short-term memory) [20].

2.4. DESCRIPCIÓN DE LAS MÉTRICAS

Las métricas utilizadas para evaluar todos los modelos es su conjunto son:

- El más bajó error en términos relativo con los datos de prueba. Se calcula como el valor verdadero sobre el valor pronosticado menos uno. La razón de calcular el error de esta forma es que la escala de todas las series es diferente.
- Que los errores tengan homocedasticidad y normalidad usando los datos de entrenamiento Breusch-Pagan [21] y Jarque-Bera [22]. Se acepta la hipótesis con un valor probabilístico mayor a 0.05.
- Tiempo promedio que toma para hiper parametrizar usando los mismos recursos computacionales. Tiempo promedio se determina como el tiempo total utilizado entre la suma total de combinaciones de los parámetros buscados.
- Para seleccionar el mejor modelo SARIMA se usó el AIC (criterio de información de Akaike) [23].

2.5. DESCRIPCIÓN DE LOS EXPERIMENTOS / SIMULACIONES

Usando 69 periodos como entrenamiento se probaron los tres diferentes modelos. En cada uno de ellos se realizó una búsqueda para encontrar el mejor modelo modificando los hiper parámetros.

6 periodos de prueba son usados para calcular el error relativo. Una vez obtenido el mejor de cada uno de los modelos, se hacen los pronósticos, se utiliza la transformación inversa y se compara con los valores reales para calcular el error relativo. También se confirma con los datos de prueba la normalidad de las innovaciones y el tiempo promedio que tardó encontrar el mejor modelo.

Para el modelo SARIMAX se busca el mejor p, d, q, P, D, Q y s.

Para el modelo SVR se busca el mejor modelo entre los parámetros gamma, epsilon, C, kernel y grado.

Para el modelo LSTM se busca entre los parámetros la profundidad de la red, la cantidad de rezagos, la cantidad de neuronas y las funciones de activación.

2.6. ENTRENAMIENTO, EVALUACIÓN, SELECCIÓN Y PRONÓSTICOS AUTOMÁTICA EN PRODUCCIÓN.

El preprocesamiento, entrenamiento, evaluación, selección y pronostico se realizará cada mes con nuevos datos, lo que lo haría un método de aprendizaje en línea (o incremental).

El razonamiento para hacer todo el proceso de nuevo cada mes es porque los datos son escasos y solo se obtienen cada mes, además que los meses inmediatos anteriores son más representativos y tienen mayor peso para los siguientes periodos.

Los pasos son:

1. Lectura o importación de datos: Proceso para importar los datos y limpiarlos de la forma requerida.
2. Detección de atípicos. Usando una metodología propuesta por el Banco de Pagos Internacionales.
    a. Estandarización de los datos y suavizamiento usando regresiones locales con el algoritmo LOWESS.
    b. Identificación de agrupamientos (clusters dynamics) con el algoritmo Affinity Propagation.
    c. Detección de atípicos en cada agrupación utilizando el algoritmo DBSCAN.
    d. Investigar y marcar como atípico con binary encoding.
3. Agrupación de series de tiempo ya definidas por geografía de mercado, por tipo de costo y por la línea de producto.
4. Tratamiento de atípicos: Fechas importantes como variables exógenas para que no tengan un peso en la generalización del modelo. Además de las fechas importantes ya mencionadas y previamente guardadas como parte del modelo que tienen efectos generalizados como el inicio de la pandemia o los cambios de tendencia, a esta lista se le incluirá los atípicos del punto anterior. Serían los atípicos en la última observación si solo se están actualizando los pesos.
5. Tratamiento del sesgo y la varianza: transformar los datos usando Box-Cox.
6. Entrenamiento o actualización: Entrenamiento e hiper parametrización de diferentes modelos o ajuste del modelo si el modelo ya fue seleccionado.
    a. Evaluación comparando modelos con métricas de error, sobreajuste o simplicidad del modelo.
7. Selección del modelo: guardar o actualizar el modelo para poder ser reutilizado.
8. Pronósticos:
    a. Realizar pronósticos.
    b. Transformar usando la inversa de Box-Cox.
9. Desagrupación de los pronósticos: Para la base de datos no tiene que estar agrupada como definida en el punto 3. Tiene que estar en el nivel de agregación más baja

posible y por eso se puede utilizar una matriz de 3 dimensiones con pesos definidos entre cero y uno (0% - 100%) para que se distribuya por cada línea de costo y línea de producto.

10. Guardar o escribir los pronósticos en la base de datos.

3. RESULTADOS Y DISCUSIÓN

Resumen: *En este capítulo se presenta los resultados obtenidos del entrenamiento y prueba de los modelos para una serie de tiempo ya que el proceso es el mismo para modelar y pronosticar para todas las demás series de tiempo de forma automática, como se menciona en el apartado 2.6.*

3.1. RESULTADOS

No se presenta gráficas puntuales para la experimentación y simulaciones, solo se usan las métricas mencionadas en la sección 2.4 para evaluar los diferentes modelos.

En la tabla 5 se muestran los valores verdaderos de la serie de tiempo analizada comparados con los resultados obtenidos con los mejores modelos.

Tabla 5

	Valor real	SARIMAX	SVR	LSTM
Ago 2022 (t+1)	$6,490,418	$5,590,305	$6,100,842	$6,362,835
Sep 2022 (t+2)	$5,868,556	$5,039,503	$6,340,737	$6,252,814
Oct 2022 (t+3)	$6,240,545	$5,887,598	$5,550,856	$6,372,801
Nov 2022 (t+4)	$4,615,811	$5,545,854	$6,021,713	$6,232,829
Dec 2022 (t+5)	$6,142,760	$6,219,239	$6,300,327	$6,385,943
Ene 2023 (t+6)	$4,747,824	$5,048,539	$5,974,135	$6,094,086

La tabla 6 muestra una comparativa de los tres modelos con las tres métricas mencionadas, incluyendo los errores de cada periodo.

Tabla 6

	SARIMAX	SVR	LSTM
Tiempo promedio	49 seg.	0.0005 seg.	3 min. 50 seg.
Normalidad	Sí	No	Sí
Homocedasticidad	Sí	No	Sí
Error en t+1	0.161	0.063	0.020
Error en t+2	0.165	-0.074	-0.061
Error en t+3	0.060	0.124	-0.020
Error en t+4	-0.168	-0.233	-0.259
Error en t+5	-0.012	-0.250	-0.380
Error en t+6	-0.060	-0.205	-0.220

La tabla 7 especifica la cantidad de combinaciones de hiper parámetros de búsqueda usados para cada modelo. Para el modelo SARIMAX se busca el mejor p, d, q, P, D, Q y s.

Para el modelo SVR se busca el mejor modelo entre los parámetros gamma, epsilon, C, kernel y grado.

Para el modelo LSTM se busca entre los parámetros la profundidad de la red, la cantidad de rezagos, la cantidad de neuronas y las funciones de activación.

Tabla 7

	SARIMAX	SVR	LSTM
Parámetros	576	91809	32

3.2. DISCUSIÓN

Recordando los resultados obtenidos en benchmark en la sección 2.2.1: En los pronósticos de uno y dos meses previos (t-1, t-2) se obtuvo un error de 3% promedio, tres meses previos (t-3) el error promedio es 6%, (t-4) 10%, (t-5) 16% y (t-6) 21%.

Los resultados en los diferentes modelos es interesante observar que el modelo SARIMAX obtiene resultados peores, comparado con los errores de referencia y los otros dos modelos, en los primeros tres periodos, mientras que en los últimos 3 fueron mejores.

Antagónicamente, los resultados del modelo de máquina de vector de soporte y la red neuronal recurrente de los primeros tres periodos fueron mejores que el modelo SARIMAX, pero no mejores que los errores de referencia. Los últimos tres periodos fueron peores que el periodo de referencia y que el modelo SARIMAX.

De los resultados se debe recordar que los pronósticos son solo tan buenos como los pronósticos de las variables exógenas, como las unidades, la base instalada o la cantidad de reparaciones. Los modelos que resultaron buenos en los tres meses se ajustan bien a las variables exógenas y por lo tanto al corto periodo de tiempo. El modelo que resultó mejor entre los cuatro a seis meses se ajusta mejor a los periodos pasados de la serie de tiempo por los componentes autorregresivos y de promedio móvil.

En cuanto a tiempo SVR es el más rápido por mucho. La red neuronal Long-Short Term Memory es las más lenta. En total la red neuronal tomo cinco horas en la hiper parametrización y SARIMAX tomó una hora.

En términos de pruebas de normalización y homocedasticidad el modelo de máquinas de vector de soporte no cumple con las pruebas.

La combinación de modelos es prometedora: Una red neuronal recurrente para los siguientes tres primeros periodos a pronosticar y el modelo SARIMAX para los siguientes tres de los seis periodos a pronosticar.

4. CONCLUSIONES

Resumen: *En este capítulo se presentan las conclusiones y trabajo futuro para la implementación de modelos de aprendizaje máquina y computación en la nube para el pronóstico de costos atribuidos a la operación de reparación de garantías.*

4.1. Conclusiones

Conclusiones prometedoras con este trabajo es la formalidad de los pronósticos en la granularidad requerida, reducción de tiempo en el proceso y limpieza de la base de datos. Se logró determinar un flujo tareas, algoritmos y pruebas a implementar para conseguir un buen pronóstico.

El uso de una red neuronal recurrente para los siguientes tres primeros periodos a pronosticar y el modelo SARIMAX para los siguientes tres de los seis periodos a pronosticar es una combinación que funciona. Sin embargo, el conocimiento que cada uno de los analistas financieros tienen para el corto plazo no es igualable con los modelos estadísticos.

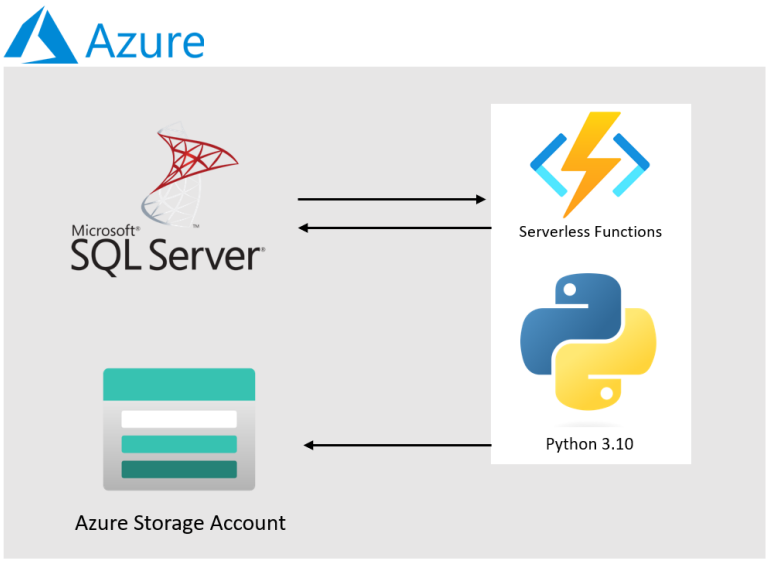
Aun se debe conseguir más datos que expliquen y pronostiquen de mejor manera, en el corto y largo plazo. Y por último reducir las irregularidades en los datos en la construcción de la base de datos.

4.2. Trabajo Futuro

Implementación en un ambiente de producción usando computación en la nube, la arquitectura es ejemplificado en la figura 6:

- Tomar los datos de la base de datos usando Python en la nube.
- Usar una función en la nube para ejecutar Python para realizar el entrenamiento y pronósticos.
- Con la misma función escribir los pronósticos en la base de datos y guardar una copia de los resultados y que permita rastrear los modelos usados.

Figura 6. Ejemplo de arquitectura en usando Microsoft Azure.



BIBLIOGRAFÍA

[1] Yves R. Sagaert, El-Houssaine Aghezzaf, Nikolaos Kourentzes, Bram Desmet. (2018). Tactical sales forecasting using a very large set of macroeconomic indicators. European Journal of Operational Research. Volumen 264. Número 2. Páginas 558-569. ISSN 0377-2217. https://doi.org/10.1016/j.ejor.2017.06.054.

[2] Wu, S. and Akbarov, A. (2011). Support vector regression for warranty claim forecasting. European Journal of Operational Research. Volumen 213. Número 1. Páginas 196-204.

[3] Spyros Makridakis, Evangelos Spiliotis, Vassilios Assimakopoulos. (2022). The M5 competition: Background, organization, and implementation. International Journal of Forecasting. Volumen 38, Número 4. Pages 1325-1336. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2021.07.007.

[4] A. David Lainder, Russell D. Wolfinger. (2022). Forecasting with gradient boosted trees: augmentation, tuning, and cross-validation strategies: Winning solution to the M5 Uncertainty competition. International Journal of Forecasting. Volumen 38. Número 4. Páginas 1426-1433. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2021.12.003.

[5] [InfoQ]. Yuan, D. (2018). *Two Effective Algorithms for Time Series Forecasting* [Video]. YouTube. https://www.youtube.com/watch?v=VYpAodcdFfA

[6] The pandas development team. (2020). pandas-dev/pandas: Pandas. Zenodo. https://doi.org/10.5281/zenodo.3509134. Software en https://pandas.pydata.org/.

[7] J. D. Hunter, (2007). Matplotlib: A 2D Graphics Environment in Computing in Science & Engineering. Volumen 9. Número 3. Páginas 90-95. 10.1109/MCSE.2007.55. Software en https://matplotlib.org/.

[8] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python**.** *Nature Methods*, Volumen 17. Número 3. Páginas 261-272. Software en https://scipy.org/.

[9] Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*. Software en https://www.statsmodels.org/.

[10] Federico Garza, Max Mergenthaler Canseco, Cristian Challú, Kin G. Olivares. (2022). StatsForecast: Lightning fast forecasting with statistical and econometric models. PyCon Salt Lake City, Utah, US 2022. https://github.com/Nixtla/statsforecast.

[11] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. Volumen 12. Número 85. Páginas 2825 – 2830. http://jmlr.org/papers/v12/pedregosa11a.html. Software en https://scikit-learn.org/stable/.

[12] Chollet, Fancois y otros. (2015). Keras. Software en https://keras.io.

[13] Abadi, Martin, Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., y otros. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software en https://www.tensorflow.org.

[14] A A Syntetos, J E Boylan & J D Croston. (2005). On the categorization of demand patterns, Journal of the Operational Research Society. Volumen 56. Número 5. Páginas 495-503. 10.1057/palgrave.jors.2601841.

[15] Liu, X., Liu, H., Guo, Q. *et al (2020).* Adaptive wavelet transform model for time series data prediction. *Soft Comput* 24, Páginas 5877–5884. https://doi.org/10.1007/s00500-019-04400-w.

[16] Benatti, N., & Alexis, A. (2021). Time series outlier detection, a data-driven approach. Bank of Italy Workshop on "Machine learning in central banking". Bank for International Settlements. https://www.bis.org/ifc/publ/ifcb57_07.pdf

[17] Box, George E. P.; Cox, D. R. (1964). "An analysis of transformations". Journal of the Royal Statistical Society, Series B. Volumne 26. Número 2. Páginas 211–252. JSTOR 2984418.

[18] Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software. Volumen 27. Número 3. Páginas 1–22. https://doi.org/10.18637/jss.v027.i03

[19] Smola, A.J., Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*. Volumen 14. Páginas 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.

[20] Chollet, F. (2017). *Deep learning with python*. Manning Publications.

[21] Breusch, T. S.; Pagan, A. R. (1979). "A Simple Test for Heteroskedasticity and Random Coefficient Variation". Econometrica. Volumen 47. Número 5. Páginas 1287–1294. 10.2307/1911963. JSTOR 1911963.

[22] Jarque, Carlos M.; Bera, Anil K. (1987). "A test for normality of observations and regression residuals". International Statistical Review. Volumen 55. Número 2. Páginas 163–172. 10.2307/1403192. JSTOR 1403192.

[23] Akaike, H. (1974), "A new look at the statistical model identification", IEEE Transactions on Automatic Control, Volumen 19. Número 6. Páginas 716–723, 10.1109/TAC.1974.1100705.

[24] Danny, V. (2020). Building your own scikit-learn Regressor-Class: LS-SVM as an example. The Delocalized Physicist. Accesado abril 2022, from https://dannyvanpoucke.be/building-scikit-learn-regressor-lssvm-en/

5. ANEXOS

5.1. DEFINICIÓN DE COSTOS

Costo	Explicación
Region Owned Expense	Costos pertenecientes a la región, propios de operaciones involucradas en reparación y asistencia.
Variable Expense	Costos variables de operaciones relacionados a la reparación, insumos o asistencia a productos.
Contact Center	Costos variables de asistencia telefónica.
Delivery	Costos variables de reparación y asistencia física.
Supply Chain	Costos variables de la cadena de suministro, insumo de partes, impuestos, logístico e inventario.
Repair OH Expense	Costos fijos o semifijos relacionados empleados de la administración y soporte de las operaciones diarias.
Contact Center OH	Costos fijos o semifijos de empleados administrativos para Contact Center.
Delivery OH	Costos fijos o semifijos de empleados administrativos para el grupo de técnicos e ingenieros.
Supply Chain OH	Costos fijos o semifijos de empleados administrativos para el grupo de cadena de suministro.
Other Warranty Expense	Otros costos regionales atribuidos a costos variables.
Worldwide Owned and Allocated Expense	Costos fijos o semifijos de empleados e inversiones globales que soportan a los tres grupos operativos de CS (Customer Support).
CS HQ	Costos fijos de empleados globales que soportan a los tres grupos operativos, incluyendo administrativos, finanzas y directivos.
CS Investments	Costos fijos de inversiones globales.
GBU Owned and Allocated	Costos de tipo diverso, fijo o variables, que incluye empleados y costos operativos de las unidades globales de negocio (Global Business Unit).
Net Reserve Expense	Reserva neta es la suma de Reserva (Accrual) más Amortización (Amortization). típicamente un número positivo.
Accrual for Shipments	Reserva de dinero que la compañía realiza con el motivo de hacer frente a sus obligaciones y pagar a sus empleados y proveedores. Basado en el costo promedio y el porcentaje de fallas esperadas del producto vendido.
Amortization	Amortización de la reserva, siempre un número negativo.