Escuela de Negocios
Tipo de documento: Tesis de maestría

Master in Management + Analytics

Desarrollo de un método automatizado para la detección de anomalías en la estructuración comercial de suscripciones

Autoría: Schmid, Carlos Maximiliano
Año: 2025

UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

DESARROLLO DE UN MÉTODO AUTOMATIZADO PARA LA DETECCIÓN DE ANOMALÍAS EN LA ESTRUCTURACIÓN COMERCIAL DE SUSCRIPCIONES

TESIS

Carlos Maximiliano Schmid

Mayo de 2025

Tutor: Damián Ilkow

Resumen

Este trabajo aborda el problema de detección de anomalías en suscripciones vendidas por una empresa que se dedica a la comercialización de software. Dado que se descubrieron riesgos en su sistema *CPQ* (*Configure, Price, Quote* – Configurar, poner Precio, Cotizar) que ya se materializaron en significativas pérdidas de facturación, se busca desarrollar una solución que permita identificar nuevos potenciales riesgos proactivamente, además de controlar los riesgos ya conocidos. Para ello, se propone un sistema que analiza cotizaciones históricas combinando técnicas de regresión y de agrupamiento no supervisado. Mediante esta solución efectivamente se logran detectar patrones anómalos conocidos y desconocidos, lo que ofrece transparencia para evaluar riesgos o beneficios estratégicos. Aunque con potencial de mejora, este enfoque fortalece la gestión de riesgos y sienta las bases para análisis futuros y decisiones de impacto estratégico.

Abstract

This work addresses the issue of anomaly detection in subscriptions sold by a company dedicated to software commercialization. Given that risks were discovered in its CPQ (Configure, Price, Quote) system, which have already resulted in significant billing losses, the aim is to develop a solution that proactively identifies new potential risks while also managing known ones. To this end, a system is proposed that analyzes historical quotes by combining regression and clustering techniques. Through this solution, both known and unknown anomalous patterns are effectively detected, providing transparency to assess risks or strategic benefits. Though improvable, this approach strengthens risk management and lays the groundwork for future analyses and decisions with strategic impact.

Índice

1 Introducción

1.1 Contexto y Problema

El problema a abordar se da en el contexto de una empresa que vende software por suscripción. Posee un sistema *CPQ* (*Configure, Price, Quote* – en español Configurar, poner Precio, Cotizar) para gestionar el proceso de configuración, cotización y generación de presupuestos para los clientes de sus soluciones ofrecidas. En más detalle:

- **Configurar (Configure)**: Permite personalizar productos o servicios según las necesidades del cliente, asegurando combinaciones válidas.
- **Poner Precio (Price)**: Se calculan automáticamente los precios, considerando descuentos, promociones o reglas específicas.
- **Cotizar (Quote)**: Generar propuestas o presupuestos para enviar al cliente, agilizando las ventas.

Este sistema se integra en el proceso de venta en las siguientes etapas:

1. **Identificación de Necesidades del Cliente**. El equipo de ventas interactúa con el cliente para entender sus requerimientos (por ejemplo, número de usuarios, módulos necesarios, nivel de soporte). En el sistema *CPQ*, se carga la oportunidad de negocio.

2. **Configuración de los Productos**. El sistema *CPQ* guía al equipo de soporte de ventas para seleccionar los productos y elegir características válidas (por ejemplo, evita combinaciones incompatibles como un plan básico con funciones premium).

3. **Cálculo de Precios y Aprobación.** El sistema calcula precios de referencia de la suscripción automáticamente, aplicando las tarifas de acuerdo a los períodos temporales correspondientes y descuentos por volumen u otras promociones. Sin embargo, los precios a ofrecer al cliente pueden ser ajustados por la fuerza de ventas. Los precios elegidos y todas las condiciones a continuación pasan por una serie de aprobadores, que se determina en base a reglas de negocio predefinidas.

4. **Generación y Envío de la Cotización**. El sistema produce una cotización detallada con el precio final, y términos y condiciones de la suscripción. La cotización luego se envía al

cliente, integrada con un sistema de gestión de relaciones con el cliente para su seguimiento.

5. **Negociación y Ajustes**. Si el cliente solicita cambios (por ejemplo, reducir la cantidad de usuarios), el vendedor ajusta la configuración en el sistema *CPQ*, el cual siempre y cuando los cambios estén en el marco de las reglas predefinidas, recalcula el precio y de volver a obtener las aprobaciones necesarias, genera una nueva cotización.

6. **Cierre, Facturación y Entrega**. Una vez que el cliente acepta la cotización, el sistema *CPQ* se integra con otros sistemas, lo que permite emitir la primera factura cuando corresponda y a desencadenar el proceso de entrega de la solución contratada.

El proceso se resume en el siguiente diagrama:

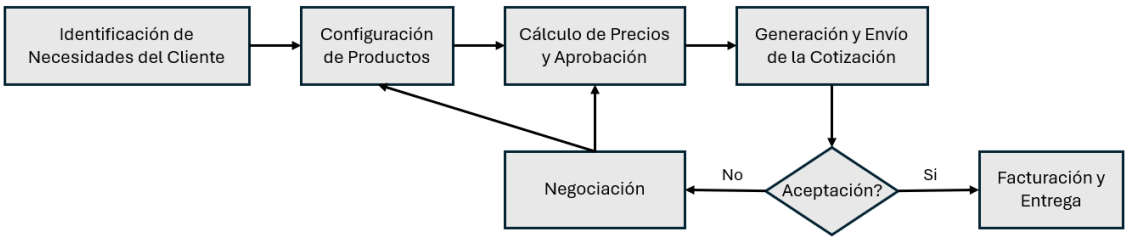

Figura 1. Proceso de venta con el sistema *CPQ.*

Unos de los elementos de esencial importancia para la salud del negocio consisten en la secuencia de aprobaciones en conjunto con las reglas de negocio que la determinan. Estos elementos tienen como principal función asegurar que los negocios propuestos sean rentables y que se adhieran a buenas prácticas, previniendo la materialización de riesgos. Si bien ya existen reglas para cubrir riesgos conocidos, tendría sentido monitorear el desarrollo del comportamiento comercial para asegurar la cobertura de nuevos riesgos que pueden surgir o que ya existan y no se haya tomado conocimiento sobre ellos.

Por ejemplo, recientemente se descubrió que la fuerza de ventas estaba aprovechando la posibilidad de manipular al sistema con ciertas estructuras comerciales de las suscripciones ofrecidas para reducir los costos que el sistema calcula y consecuentemente los precios que se pueden ofrecer, lo que constituye un riesgo de disminución de la rentabilidad o incluso de pérdidas. En unos pocos casos donde este riesgo se materializó de forma extrema, el impacto negativo en términos de facturación superó el millón de dólares a lo largo de toda la duración de la suscripción de un cliente. También hay casos más sutiles, en los que el impacto se estima

en cientos o decenas de miles. Es difícil estimar un valor a nivel agregado, pero éste definitivamente es lo suficientemente grande como para que se justifique un análisis más detallado del problema.

Para cubrir este nuevo riesgo que se descubrió, se desarrollaron nuevas reglas de negocio que fueron implementadas en el sistema *CPQ*, el cual incluye aprobadores adicionales ante la detección de este comportamiento y alerta a los aprobadores. Cabe destacar que hasta que el sector de Control de Gestión detectó este comportamiento, pasaron meses durante los cuales esta vulnerabilidad del sistema fue aprovechada. Y la detección fue por comunicación informal, es decir, no en base a revisiones regulares, ni de alertas en algún sistema. Se trataba de casos en los que no se sabía ni siquiera qué había que verificar. Se trataba de riesgos desconocidos. No se sabía que existían, por lo que no había controles. Una vez que se tomó conocimiento de la situación, se hizo un análisis a medida para ese riesgo en particular que se había descubierto. Este análisis fue en tablas de cálculo en base a datos históricos. En la sección 2.2 (Variables Disponibles y Preprocesamiento Inicial) de este trabajo, se ampliará en cómo se hizo ese análisis, ya que parte del mismo se replicará aquí para etiquetar datos.

Por ocurrencias como estas, se genera la necesidad de contar con un sistema que detecte de forma automática casos como estos o con nuevos comportamientos riesgosos para el negocio, de modo que se puedan establecer nuevas reglas para limitar esos nuevos riesgos. Es decir, existe la necesidad de identificar o detectar nuevos riesgos, además de controlar los ya conocidos, para lo cual actualmente no hay un proceso estandarizado.

1.2 Objetivo

Como se presume que los casos riesgosos son poco frecuentes, el objetivo de este trabajo es desarrollar un sistema de detección de anomalías que facilite su identificación temprana y permita reducir el tiempo de reacción del área de Control de Gestión para tomar las medidas apropiadas.

Esto sería de especial ayuda teniendo en cuenta que la forma de trabajo actual tiende a ser más reactiva, es decir, se implementan nuevas medidas o reglas una vez que un comportamiento riesgoso ya pasó a adoptarse en mayor medida y hacerse conocido, de modo que ya pudo haber tenido un impacto negativo a mayor escala. Con un método automatizado, los comportamientos anómalos se pueden detectar tempranamente y su nivel de riesgo puede evaluarse para determinar si hace falta implementar nuevas medidas o si, en el caso que la anomalía no constituya un riesgo, puede ignorarse. Esta determinación de riesgo se haría con la mirada de un analista. Este trabajo se enfoca en la identificación de los casos, para luego poder ponerlos a

disposición del análisis de riesgo mencionado. Cabe hacer énfasis en que no todo comportamiento anómalo necesariamente es riesgoso. Tampoco puede asegurarse que todo comportamiento riesgoso sea anómalo, pero el hecho de detectar anomalías es un camino posible para detectar comportamientos riesgosos u oportunidades de mejora en los procesos del negocio.

1.3 Revisión de Bibliografía

Se realizó una revisión de bibliografía sobre el tema y a la fecha no se encontró contenido público específicamente para la detección de anomalías en el contexto de la estructuración comercial de suscripciones. Se presume que esto se debe a que la temática aborda cuestiones muy específicas al modelo comercial de cada compañía, de forma que no constituye un problema que suele llega al ámbito público.

Sí se encontró bibliografía referida a la detección de anomalías aplicada problemas más genéricos con prevalencia en ámbitos como la detección de fraude financiero, ciberseguridad, sistemas industriales, control de calidad, comportamiento del usuario en comercio electrónico y salud.[1] [2] El hecho de que cada vez haya más datos[3] permite que los mismos puedan aprovecharse para más finalidades, siendo una de ellas la detección de anomalías. En este contexto de *big data*[4] uno de los principales desafíos para la detección de anomalías es el problema de la alta dimensionalidad, en el que la complejidad del análisis de datos se incrementa con respecto a la cantidad de dimensiones, requiriendo métodos de procesamiento más sofisticados[5]. Éste es un problema estudiado por Thudumu, S., Branch, P., Jin, J., y Singh, J. (2020). Tradicionalmente hay gran variedad de métodos estadísticos para la detección de anomalías, y muchos de ellos también pueden utilizarse con ciertas consideraciones en al ámbito

[1] Ersoy, P. (2023). Anomaly detection: Identifying critical issues to reduce revenue losses. Dataroid. https://www.dataroid.com/post/anomaly-detection-identifying-critical-issues-to-reduce-revenue-losses

[2] Bollu S. (2024). Anomaly Detection of User Behavioral Events in E-commerce Electronics Stores using SVMs. Blekinge Institute of Technology. https://www.diva-portal.org/smash/get/diva2:1888826/FULLTEXT01.pdf

[3] Taylor, P. (2024). *Worldwide data created*. Statista. https://www.statista.com/statistics/871513/worldwide-data-created/

[4] Se entiende por *big data* a la gran cantidad de datos que son generados por personas y máquinas diariamente, de la cual se puede extraer información. Por su gran volumen, requiere técnicas especiales de procesamiento.

[5] Thudumu, S., Branch, P., Jin, J., & Singh, J. (2020). A comprehensive survey of anomaly detection techniques for high dimensional big data. Journal of Big Data, 7(1), Article 42. https://doi.org/10.1186/s40537-020-00320-x

de *big data.* Por ejemplo, argumentan que uno de los enfoques más simples para manejar este problema es minimizar la cantidad de variables a un subconjunto de todas las dimensiones disponibles, lo que puede ayudar a identificar anomalías que estarían ocultas si uno analizara todo el conjunto de datos en su totalidad. Otra posibilidad es utilizar técnicas de reducción de dimensionalidad como Análisis de Componentes Principales o Escalamiento Multidimensional, entre otras.

Otras técnicas son las basadas en aprendizaje no supervisado y técnicas de agrupamiento, en las cuales se puede asumir que las observaciones normales son parte del mismo clúster, mientras que las anómalas tenderían a estar en otro clúster. Por otra parte, Natha, S., Leghari, M., Rajput, M. A., Zia, S. S., & Shabir, J. (2024)[6] mencionan que se pueden considerar observaciones como anómalas si además se encuentran lejos del centroide de cada clúster. Por otro lado, Ertoz et al. introdujeron un método de vecino compartido más cercano que puede manejar la multidimensionalidad.

Por otra parte, hay técnicas basadas en densidad, que identifican zonas mayor y menormente pobladas del espacio de datos. Thudumu, S., Branch, P., Jin, J., & Singh, J. (2020) explican que estas tienden a no ser efectivas cuando la dimensionalidad de los datos aumenta. Sin embargo, Chen et al. desarrollaron un estimador de densidad que funciona con buen desempeño y precisión en conjuntos de datos multidimensionales.

Las mencionadas son solo algunas, pero existe gran cantidad de investigación realizada sobre muchas otras técnicas de Aprendizaje Automático y Aprendizaje Profundo para la detección de anomalías de forma supervisada, semi supervisada y no supervisada. Como naturalmente cada técnica tiene sus ventajas y desventajas, también se suelen combinar técnicas en enfoques de ensamble para superar las desventajas de los métodos basados en una única técnica, y mejorar la resiliencia y precisión del sistema en su conjunto. [7]

Otro desafío que aplica a la detección de anomalías en general es el de la escasez de datos etiquetados, ya que estos no siempre están disponibles. De ser así, la posibilidad de usar modelos de entrenamiento supervisado queda fuertemente limitada.[8] Este es el caso del

[6] Natha, S., Leghari, M., Rajput, M. A., Zia, S. S., & Shabir, J. (2024). A Systematic Review of Anomaly detection using Machine and Deep Learning Techniques. Queueing Systems.

[7] Ramasamy, V. U. (2024). Overview of anomaly detection techniques across different domains: A systematic review. International Journal of Computer and Experimental Science and Engineering.

[8] Bablu, Tarek & Mirzaei, Hossein. (2025). Machine Learning for Anomaly Detection: A Review of Techniques and Applications in Various Domains.

presente trabajo, ya que si bien se cuentan con algunos datos etiquetados, se apunta a descubrir anomalías de nuevos tipos, lo que genera la necesidad de utilizar técnicas de aprendizaje semi supervisado o no supervisado.

Además, la escasez de datos etiquetados se ve exacerbada por otro de los desafíos típicos de la detección de anomalías, que es el de los datos desbalanceados por naturaleza. Bablu, T. & Mirzaei, H. (2025) mencionan que el hecho de que las anomalías sean una pequeña porción de todos los datos, le puede dificultar a los algoritmos aprender las características de las anomalías, ya que se ven eclipsadas por las clases mayoritarias.

Dejando lo técnico de lado brevemente y pasando a algunos datos de contexto: El mercado de detección de anomalías está en aumento y se prevé que siga creciendo en los próximos años. Por ejemplo, un reporte de *Market Research Future*[9] menciona que el tamaño del mercado combinado de Norteamérica, Sudamérica, Europa, Asia Pacífico, Oriente Medio y África (representando los ingresos totales generados con ventas de productos y servicios de detección de anomalías) se estimó en 2,56 mil millones de dólares en 2023 y en 2,88 en 2024. Se espera que su tasa de crecimiento anual compuesta sea de 12,48% entre 2025 y 2035. Este crecimiento viene dado por un aumento en la demanda de este tipo de soluciones, que asimismo se origina en un creciente reconocimiento sobre la importancia de tener sistemas de detección de anomalías para la optimización de las operaciones de las compañías. Esto se da en entornos con crecientes riesgos de negocio e informáticos, y con rápidos avances tecnológicos, especialmente en el área de inteligencia artificial, que si bien generan nuevos riesgos, también permiten el desarrollo de mejores sistemas de detección. Por otra parte, para tomar dimensión del problema de fraude solamente, se estimó que las pérdidas a nivel global fueron de 5,127 billones de dólares en 2019, lo que representa alrededor de 6% del Producto Bruto Interno mundial.[10]

En resumen, si bien ya hay gran cantidad de investigación realizada sobre detección de anomalías en ámbitos varios, los cuales cada vez son más, para la estructuración comercial de suscripciones no se encontraron trabajos públicos existentes. Por este motivo se genera la necesidad de desarrollar un enfoque que permita aplicar técnicas existentes a este dominio específico, con las adaptaciones que sean necesarias. La aplicación de métodos existentes a un problema complejo de un ámbito específico como el mencionado es uno de los principales

[9] Dhapte, A. (2025). Anomaly detection market size, growth report - 2030. Market Research Future. https://www.marketresearchfuture.com/reports/anomaly-detection-market-5756

[10] Crowe, & Centre for Counter Fraud Studies. (2019). *The financial cost of fraud 2019*. https://www.crowe.com/global/news/fraud-costs-the-global-economy-over-us$5-trillion

aportes de valor de este trabajo. En el mismo, se desarrollará un método de ensamble, comprendido principalmente por una etapa que utiliza técnicas basadas en distancia sobre un subconjunto de los datos, y luego una etapa de agrupamiento no supervisado para segmentar los resultados. Un desafío adicional que se va a tener que superar en este trabajo es el tratamiento de datos secuenciados con longitud variable, problema para el cual no se encontró bibliografía que lo aborde, por lo menos con datos de características similares a los que se tienen. En la siguiente sección se explicarán más detalles sobre la naturaleza de los datos.

2 Datos

2.1 Información General y Definición de la Unidad de Análisis

Se cuenta con un conjunto de datos que contiene todos los ítems de las cotizaciones de ventas iniciales (no ventas adicionales sobre contratos preexistentes ni renovaciones) que llevaron a la firma de un contrato desde Enero de 2022 hasta Septiembre de 2024, y consecuentemente a su entrada en vigencia. Es por este motivo que a lo largo del trabajo se utilizarán las palabras "cotización" y "contrato" con el mismo sentido a fines prácticos.

Es importante tener en cuenta que dentro de una cotización, puede haber distintas soluciones (identificables mediante *SKUs – Stock Keeping Units,* Unidades de Seguimiento de Stock en español, que son códigos de identificación unívocos de una solución o producto), y de cada *SKU* dentro de la cotización puede haber distintas fases con distintas duraciones y cantidades (ejemplo en la tabla de la figura 2). Nótese que la cotización 1 tiene los *SKUs* A y B. Asimismo el *SKU* B dentro de la cotización 1 tiene dos fases. Una con 15 y otra con 25 unidades. Además, cada fase puede tener una duración distinta. Cómo se combinan estas variables (duración, cantidad, valor) es lo que llamaremos la estructura comercial de un contrato o de una cotización.[11]

Cabe aclarar que dentro de cada fase puede haber cortes temporales por motivos administrativos que generan distintos ítems en el conjunto de datos, es decir que una fase puede estar compuesta por varios ítems. Un ejemplo de esto se puede ver en la figura 2 en el *SKU* "B" de la Quote 3 (naranja pálido). Nótese que sus últimos dos ítems tienen la misma cantidad y conforman una única fase de esa cotización.

Vamos a definir la unidad de análisis como una secuencia de longitud variable, compuesta por fases de un mismo *SKU* dentro de una cotización. En la figura 2 cada unidad de análisis tiene un

[11] DealHub. (n.d.). *Deal structure*. DealHub. https://dealhub.io/glossary/deal-structure/

color distinto para facilitar la representación gráfica de este concepto. En definitiva, una unidad de análisis está conformada por todos los ítems secuenciados de un par Cotización-*SKU*.

Cotización	SKU	Fecha de Inicio	Fecha de Finalización	Cantidad
1	A	01-01-24	31-12-24	3
1	B	01-01-24	30-06-24	15
1	B	01-07-24	31-12-24	25
2	G	01-01-23	30-06-25	7
2	G	01-07-25	31-12-25	12
3	A	01-01-24	31-12-25	9
3	B	01-01-24	30-06-24	30
3	B	01-07-24	31-12-24	60
3	B	01-01-25	31-12-25	60
3	F	01-01-24	31-12-25	9
3	T	01-01-24	30-06-24	55
3	T	01-07-24	31-12-25	70

Figura 2. Estructura del conjunto de datos con sus variables clave.

El comportamiento de cómo se estructuran las fases es uno de los principales aspectos a estudiar, ya que dependiendo de las cantidades, se permiten distintos niveles de precios. Por ejemplo, en la figura 3 puede verse como el precio de un *SKU* disminuye a medida que la cantidad suscripta aumenta. Esto en general se debe a efectos de economía de escala. Mientras mayor el volumen, menor el costo unitario y consecuentemente menor el precio que se puede ofrecer.

Precio por Unidad	Niveles de Volumen	
	Desde	Hasta
748,00 $	1	19
480,00 $	20	49
383,00 $	50	99
305,00 $	100	199
239,00 $	200	499
192,00 $	500	Infinito

Figura 3. Ejemplificación de la influencia del volumen sobre el precio unitario.

También es importante remarcar que el sistema *CPQ* actual aplica a todas las fases de un *SKU* dentro de una cotización el precio correspondiente al volumen más alto alcanzado. Es decir, para el *SKU* de la figura 3, si se vende una suscripción de tres años, de los cuales los primeros dos son con una unidad, y el último con quinientas unidades, el precio de lista total cotizado para cada año se calcula como:

Año 1: 1 x 192 $ = 192 $

Año 2: 1 x 192 $ = 192 $

Año 3: 500 * 192 $ = 96.000 $

Puede notarse que en los años 1 y 2 se obtiene una muy significativa reducción de precio (de 748 $ a 192 $). Este ejemplo muestra un caso teórico extremo, pero sirve para explicar cómo funciona el sistema y demuestra una de sus vulnerabilidades, ya que los precios de cada nivel o escalón de volumen en realidad están basados en su correspondiente nivel de costos, no en el del más alto alcanzado. El sobre aprovechamiento de esta forma en la que funciona el sistema puede llevar a casos de pérdidas (o al menos fuerte reducción de facturación) y actualmente se controla mediante reglas de negocio implementadas en el sistema *CPQ* alertando a los aprobadores de la cotización, tal como se había mencionado en la introducción.

A la estructura comercial del ejemplo de arriba se la denomina *Hockey Stick* o palo de hockey debido a su larga duración con poco volumen y un incremento notorio hacia el final de la suscripción con una corta duración. Podemos ver otro ejemplo de forma gráfica en la figura 4.

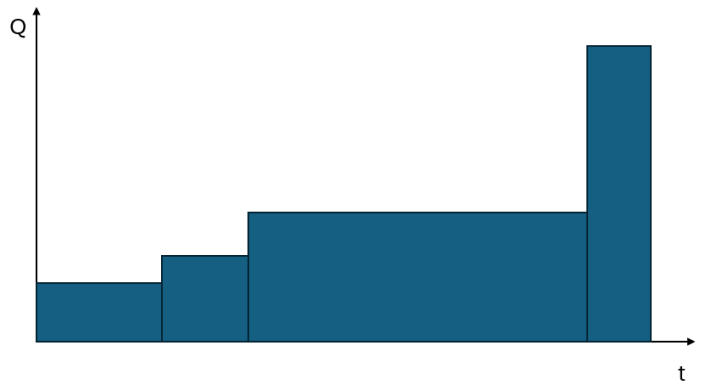

Figura 4. Representación gráfica de un contrato *Hockey Stick*

Teniendo en cuenta el impacto negativo que esta estructura puede tener, uno podría preguntarse, ¿por qué no modificar este modelo de Pricing, de modo que en cada fase el precio se calcule en base a su correspondiente Nivel de volumen? Hay varios motivos por los que se decidió no hacerlo:

- Mantener un único precio por unidad a lo largo de toda la duración del contrato simplificando la estructura contractual.
- Incentivar la contratación de cantidades superiores (mediante precios menores por volumen mayor – pero sin abusos) para fomentar la adopción de las soluciones vendidas y aumentar la facturación.

- El cambio implicaría una modificación de gran escala en los sistemas internos, que se preferiría evitar.

Es por estos motivos que el actual modelo de precios se considera una constante y se busca monitorear su uso para prevenir el aprovechamiento de riesgos no controlados y controlar los riesgos conocidos.

2.2 Variables Disponibles y Preprocesamiento Inicial

Habiendo comprendido las implicancias del apartado anterior, volvamos al conjunto de datos. Si bien en la figura 2 ya se ven algunas variables clave, a continuación se describen todas las variables con las que se cuenta:

Enteros:

- Identificadores varios, entre ellos, número de cotización, cliente, *SKU*
- Cantidad suscripta
- Nivel de volumen (Del ítem)
- Nivel de volumen máximo (de la secuencia de ítems de un *SKU* dentro de un contrato – agregada manualmente)
- Duración del ítem (días)
- Duración del contrato (suma de las duraciones de los ítems de un *SKU* – agregada manualmente)
- Duración del nivel de volumen (suma de las duraciones de los ítems de un *SKU* que están en un mismo nivel – agregada manualmente)

Punto Flotante:

- Valor Neto ($)
- Distintos puntos de precio usados para incentivos de ventas (precio de lista, precio piso, entre otros.)
- Descuento porcentual
- Costos estimados por el sistema *CPQ* ($)
- Margen estimado por el sistema *CPQ* (%)
- Duración de la última fase (%)

Fechas (desdobladas en enteros):

- Inicio y fin de suscripción
- Creación de la cotización

- Firma/Aceptación del contrato

Booleanos:

- Cálculo de costos específico para el ítem y cliente en cuestión (indica el reemplazo del cálculo realizado por defecto de forma automática)
- Ítem correspondiente al nivel de volumen más alto de la secuencia (agregado manualmente)
- Contrato *Hockey Stick* (etiquetado de forma manual)

Texto/Categóricas:

- Nombre de la solución o del *SKU*
- Divisa

Por motivos de confidencialidad, varias de estas variables fueron modificadas, por ejemplo escalándolas. Esto se hizo de una forma que no afecta el análisis en cuestión.

Además, se eliminaron ítems con valor monetario neto igual a cero, que son utilizados para fines internos, es decir, no para cotizaciones o contratos con clientes.

En el caso de la variable *Hockey Stick*, ésta se generó con un etiquetado manual, similarmente a cómo se había hecho el análisis cuando este tipo de casos se había descubierto por primera vez: Se buscaron contratos con un salto a un nivel de volumen superior en su última fase, siempre y cuando la duración de esa última fase (incluyendo todos sus ítems) fuera menor al 10% de la duración total del contrato. Este criterio de corte se definió de esta forma con la intención de etiquetar los casos más extremos, es decir en los cuales la última fase con salto de nivel de volumen tiene la menor duración relativa al total de la suscripción. Al mismo tiempo, con este criterio, se alcanzaba una cantidad suficiente de casos como para poder guiar el desarrollo de la solución. Si se elegía un corte menor al 10%, la cantidad de casos etiquetados tendía a ser insuficiente, mientras que si se aumentaba el límite, se incluirían casos con un impacto que a priori no parece tan severo. En resumen, se consideró que el 10% de la duración proveía un corte con un balance adecuado entre ambos aspectos, aunque vale aclarar que no se descarta la presencia de algunos *inliers* en las secuencias restantes.

Para poder realizar este etiquetado, se tuvieron que crear varias variables adicionales: Primero hubo que identificar para cada secuencia de *SKUs* dentro de un contrato, cuál es su nivel de volumen máximo, cuál es su duración total, y cuál es la duración de todos los ítems en el nivel de volumen máximo, es decir de la fase con mayor cantidad suscripta. Con estos datos se pudo

calcular cual es el porcentaje de duración del nivel de volumen superior, respecto a la duración total del contrato. Y con esta variable se pudo etiquetar a los contratos *Hockey Stick*.

En cuanto a las fechas, se utilizaron para luego generar algunas variables más, en concreto, se obtuvo el día de semana y el número de trimestre. Esta información se utilizará a continuación en la sección 2.3.

2.3 Análisis Descriptivo

En base a los datos con su preprocesamiento inicial, se realizó un análisis descriptivo para comenzar a comprenderlos mejor y ver si ya se pueden descubrir patrones útiles para el problema a afrontar.

Se cuenta con 116.279 ítems correspondientes a 25.747 cotizaciones. De ellos, 354 ítems correspondientes a 45 cotizaciones fueron etiquetados manualmente como manipulados en la variable *Hockey Stick*. En términos de ítems, representan 0,3% de los casos, y en términos de cotizaciones, 0,17%. Esto es un primer indicador de que el tipo de casos que se desea encontrar es muy infrecuente en todo el conjunto de datos, lo que implica un alto grado de desbalance en los datos. Este problema desemboca justamente en una de las cuestiones núcleo que se desean resolver en este trabajo. Es decir, se desea encontrar estos casos poco frecuentes con el menor esfuerzo manual posible. Esta situación luego será tomada en cuenta como un factor a considerar en el desarrollo de la solución propuesta (sección 3).

En la figura 5 se puede encontrar una matriz de correlaciones con la lista actual de variables numéricas (incluyendo el preprocesamiento completo). La matriz muestra que hay muchas variables con poca correlación. Las que muestran niveles de correlación más marcados eran esperables de antemano, por ejemplo, fechas posteriores con identificadores de mayor numeración, mayores costos de obtención de contrato (*Loss Making Floor Price*) a mayores costos de cumplimiento de contrato (*Fulfillment Costs*), o un umbral que se utiliza para aumentar los incentivos de la fuerza de ventas para que venda a precios superiores (*Low End Price)* mayor con mayores precios piso (*Floor Price*). En resumen, si bien esta matriz de correlaciones ayuda a comprender la relación entre las distintas variables, no aporta información valiosa para nuestro problema.

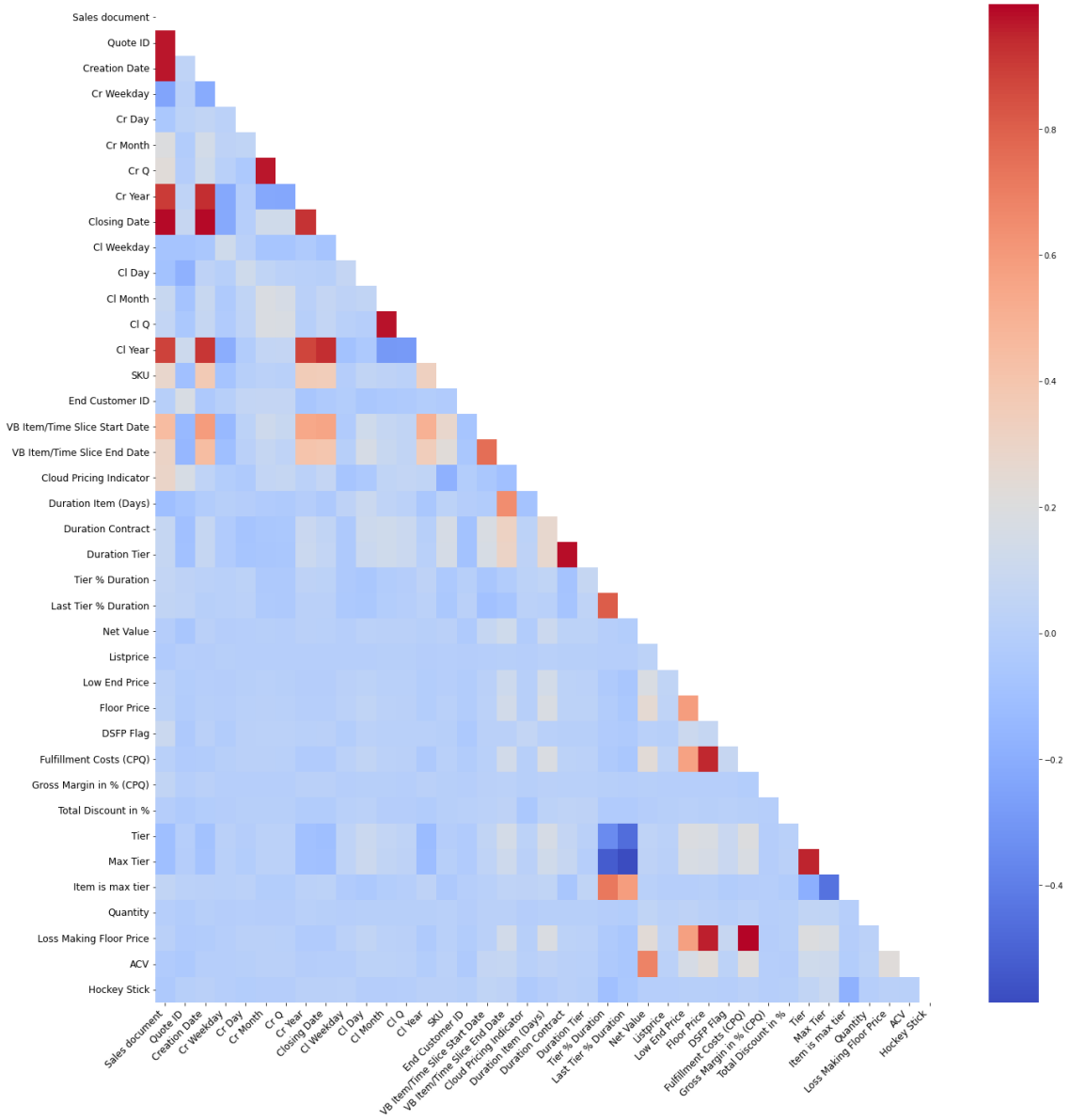

Figura 5. Matriz de Correlaciones de las variables disponibles.

Luego pasamos a explorar los datos basados en fechas. En el histograma de la figura 6 se puede ver que la mayoría de los ítems tienen una duración de un año y que en general los múltiplos de 365 días son más frecuentes que otras duraciones. Se destaca también la gran cantidad de ítems con menos de 100 días de duración.

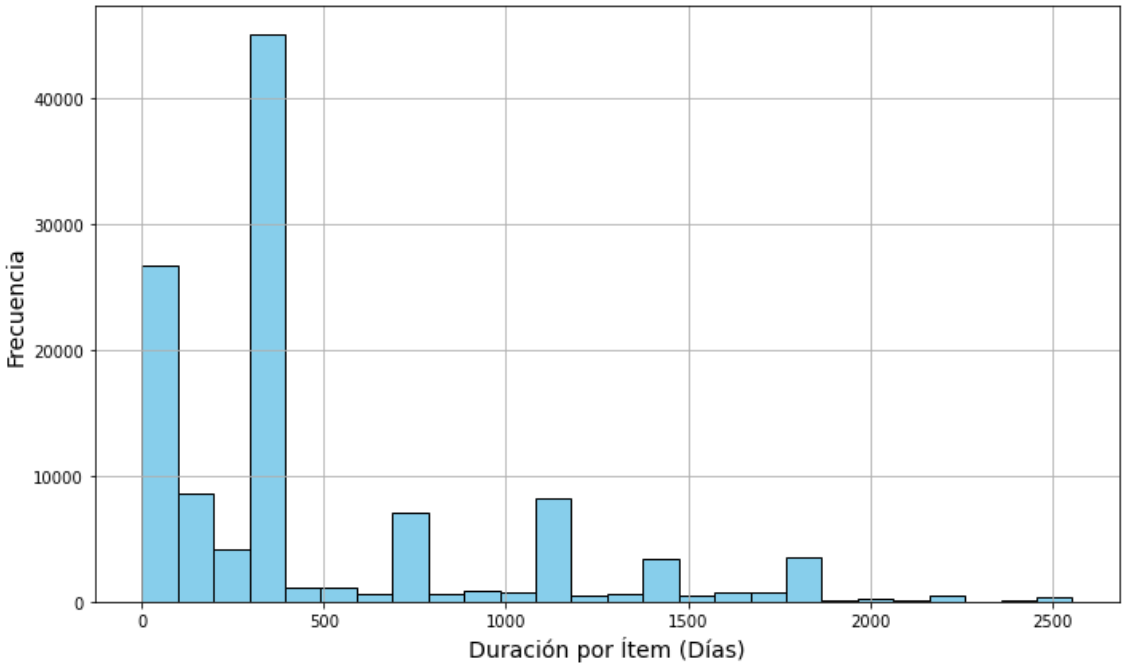

Figura 6. Distribución de frecuencias de la duración de los ítems.

Por otra parte, en la figura 7 comparamos entre el total de los datos (izquierda) y los datos marcados por la variable *Hockey Stick* (*HS* - derecha) las fechas de cierre (firma) de los contratos/cotizaciones en términos de cantidad de ítems por mes. En el total, se nota una fuerte concentración al final de cada trimestre y un crecimiento a lo largo del año, mientras que en los *HS* puede verse una concentración principalmente a mitad de año, aunque diciembre sigue siendo el mes con más ocurrencias, probablemente debido a los esfuerzos para cumplir objetivos a fin de año. Por el momento, no podemos sacar más conclusiones de esto, pero sí vale la pena tenerlo en cuenta.

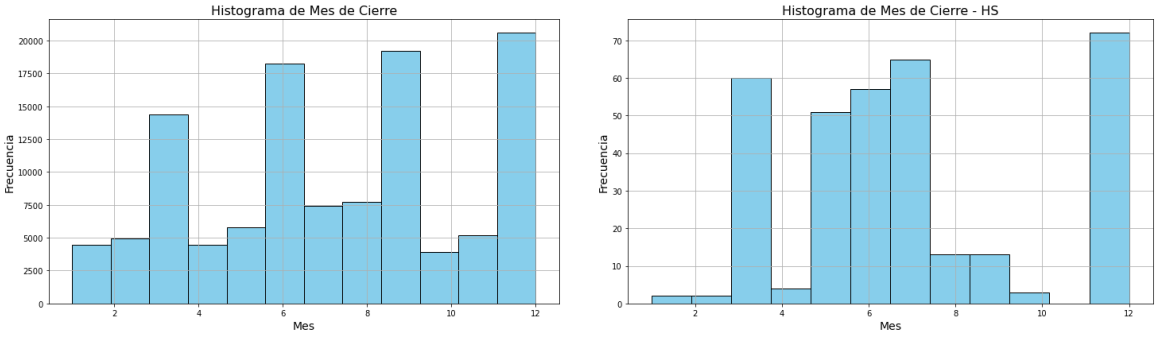

Figura 7. Contratos cerrados por mes: Total y HS.

Hacemos lo mismo con la distribución de frecuencias de *SKUs* que puede haber en una cotización (figura 8) y vemos que la mayoría tiene menos de cinco. Un *SKU* por cotización es lo más común. Esta distribución parece ser muy similar en cotizaciones *HS*, aunque en el total, si bien poco frecuentes, hay algunas cotizaciones con mayor cantidad de *SKUs*.

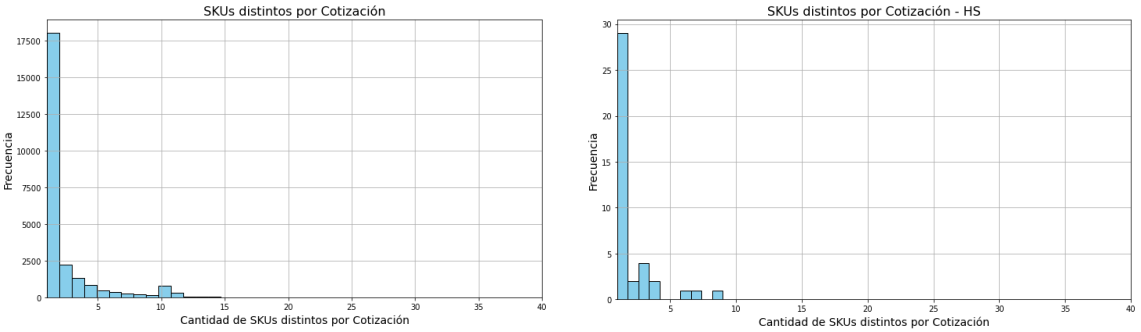

Figura 8. *SKUs* por cotización: Total y *HS*.

De forma análoga, en la figura 9 observamos la cantidad de ítems de un *SKU* dentro de una cotización, viendo que en el total lo más común es que cada *SKU* tenga solo un ítem. Esto puede ser importante ya que se estima que los casos de mayor interés para estudiar son los que tienen varios ítems/fases por *SKU*. De hecho, en el grafico *HS* de la derecha, la distribución se mueve a valores superiores a 1, siendo el valor más común el tres y con fuerte presencia también de frecuencias de entre cuatro y seis.

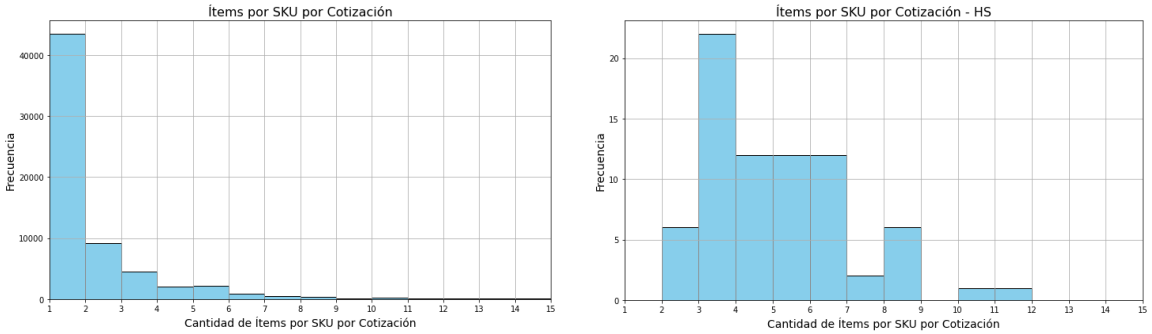

Figura 9. Ítems de un *SKU* dentro de una cotización: Total y *HS*.

Por otra parte, en la figura 10 se compara la media de los desvíos estándar intra-cotización de la duración de los ítems. La comparación se hace entre cotizaciones que tienen por lo menos una secuencia etiquetada como *Hockey Stick* contra las que no tienen ningún ítem marcado. Puede verse que los casos *Hockey Stick* tienen un desvío mucho mayor, lo que implica una dispersión de las duraciones de sus ítems también mucho mayor.

Hockey Stick	Falso	Verdadero
Duración: Desvío Estándar Promedio Intra-Cotización (días)	144,24	359,75

Figura 10. Comparación de desvío estándar de duración de ítems.

Resumiendo este análisis descriptivo, pudimos obtener cierto conocimiento acerca del comportamiento de algunas variables y pudimos ver que en el caso de las frecuencias de ítems para cada *SKU* por cotización hay una diferencia notable en su comportamiento comparando el total de los datos con los casos que se etiquetaron como del tipo *Hockey Stick*. También pudimos ver que el desvío estándar medio intra-cotización de las duraciones de ítems es notablemente mayor en los casos *Hockey Stick.* Siendo estas las únicas diferencias que se notaron hasta ahora, se debe desarrollar una solución de reconocimiento de otra manera.

3 Metodología e Implementación

Recordemos que el objetivo del trabajo es desarrollar un método para detectar anomalías que puedan tener un riesgo para el negocio, y que este método reduzca los esfuerzos de procesamiento manual en la mayor medida posible. Además, ya tenemos algunos casos riesgosos etiquetados manualmente. Se espera que el método a desarrollar detecte esos casos y que además de ellos también encuentre otros tipos de anomalías desconocidas aún.

Si solo se quisieran encontrar más casos del tipo *Hockey Stick*, podríamos entrenar un modelo supervisado con los datos etiquetados, o incluso un algoritmo más simple que identifique secuencias de *SKUs* (unidad de análisis definida en la sección 2.1) que tengan una última fase de corta duración y que muestren un salto a un nivel de volumen superior al de la fase anterior. En este caso, también sería pertinente utilizar alguna estrategia de generación de datos sintéticos para facilitar el entrenamiento del modelo supervisado con más instancias de la clase *Hockey Stick.* Sin embargo, en este trabajo se desea ir más allá de eso. Además de encontrar ese tipo de casos ya conocidos, se busca identificar de forma proactiva otros tipos de anomalías desconocidas. Teniendo en cuenta este objetivo, se terminará utilizando un modelo no supervisado, y si se generaran datos sintéticos sólo para la clase *Hockey Stick* y no para otros tipos de anomalías que aún no se conocen, se podría estar representando esta clase por demás, lo que generaría un sesgo adicional.[12] Es por esto, que es necesario utilizar otro método que pueda tratar por igual cualquier tipo de anomalía, incluso sin conocerlo de antemano.

[12] De hecho, Goldstein y Uchida (2016) evalúan distintos tipos de algoritmos no supervisados para la detección de anomalías, y en los pasos de preprocesamiento no hay referencia alguna a la generación de datos sintéticos, implicando que esta no es una práctica común o necesaria en modelos no supervisados.

3.1 Criterios de Evaluación de Éxito

Para guiar y evaluar el éxito de la solución propuesta, se definen los siguientes criterios:

1. **Facilidad para detectar anomalías de tipos ya conocidos (*Hockey Stick)*:** Se espera que éstas puedan ser detectadas con menor esfuerzo a cómo se había hecho en el análisis a medida. En ese análisis se habían procesado datos similares a los utilizados en este trabajo mediante la generación manual de variables en planillas de cálculo. Este proceso tomó horas de trabajo, especialmente considerando que era la primera vez que se había hecho, teniendo que desarrollar el método desde cero. Si bien repetir ese proceso ahora sería más rápido que desarrollarlo desde cero, el procesamiento y análisis manual en las tablas de cálculo seguiría consumiendo horas de trabajo que podrían ser ahorradas de existir una solución que realice este procesamiento de forma automatizada.
2. **Capacidad de detectar tipos de anomalías desconocidos hasta el momento:** Se espera que la solución propuesta pueda detectar nuevos tipos de anomalías, para lo que actualmente no hay una solución automatizada, ni proceso alguno implementado. Esto permitiría que la solución puede ser utilizada como una herramienta de monitoreo flexible, que pueda alertar sobre comportamientos variados sin depender de que un analista tenga que pensar cuáles podrían ser posibles comportamientos anómalos y tener que diseñar criterios específicos para cada tipo de caso para poder verificar su existencia o ausencia. Este diseño e implementación de reglas fijas consumiría demasiado tiempo y sería poco flexible sin intervención manual. Por estos motivos no hay un proceso para lograr este objetivo actualmente y se espera una solución flexible que sin reglas fijas pueda detectar distintos tipos de anomalías automáticamente.
3. **Injerencia de falsos positivos y falsos negativos:** Se espera que la solución propuesta reduzca al mínimo el tiempo que un analista deba invertir en distinguir casos que realmente son anómalos de casos que la solución sugiere que son anómalos, aunque en realidad no lo sean. Idealmente se debería poder llegar a una segmentación que pueda agrupar claramente distintos tipos de anomalías, distinguiéndolas de casos normales. Todos los casos asignados a algún grupo de anomalía efectivamente deberían tener un comportamiento que llame la atención por no ajustarse a reglas de negocio o buenas prácticas. En estos grupos, no debería haber casos que tengan comportamientos que estén alineados con buenas prácticas del negocio. Análogamente, los grupos a los cuales se les asigne casos normales, no deberían contener casos con comportamientos extraordinarios llamativos. Qué tan bien la solución pueda separar precisamente los

casos automáticamente afecta directamente la cantidad de horas adicionales que un analista debe invertir en procesar y analizar los resultados.

En la sección 4.1 (Desempeño de la Solución) se entrará en mayor detalle a la hora de aplicar estos criterios a la siguiente solución propuesta.

3.2 Primera Etapa: Regresión

3.2.1 Modelo e Intuición

Como se presume que los casos anómalos son justamente infrecuentes (ver sección 2.3), se podría tratar de comprender cual es comportamiento típico de cada *SKU* en base a todos los datos disponibles para luego identificar los casos que más se alejan de ese comportamiento. Y por comportamiento en este caso nos referimos a cómo evoluciona la cantidad suscripta en una secuencia de fases de un *SKU*. Es decir, se podría tratar de entender para cada *SKU*, qué cantidad suele tener suscripta en distintos momentos de un contrato.

Si bien idealmente este paso se debería efectuar únicamente en base a secuencias normales, en un estado inicial no hay forma de segregar todas las secuencias anómalas con certeza. A pesar de que se pudieron identificar algunas secuencias del tipo *Hockey Stick*, estas no necesariamente son todas, y además puede haber otros tipos de anomalías que no se conocen aún. Teniendo en cuenta esto y la presunción de que las anomalías son infrecuentes, para ser consistente con cualquier tipo de anomalía, se decide utilizar la totalidad de los datos para esta etapa, asumiendo el riesgo de poder estar introduciendo cierto sesgo.

A continuación se describe el método propuesto. El mismo sólo utiliza las siguientes variables de todas las disponibles en el conjunto de datos:

- SKU
- Cotización
- Cantidad Suscripta
- Fechas de Inicio y Fin de Suscripción
- Duración del Ítem

Si bien las demás variables no son utilizadas por el modelo propuesto, sí pueden llegar a ser útiles a la hora de interpretar y seguir analizando los resultados, además del fin descriptivo que ya vimos. Se aclara también, que la variable *Hockey Stick,* que identifica los casos etiquetados manualmente, es una de las que no es utilizada por el modelo ya que se utilizará un modelo no supervisado, pero sí se utiliza posteriormente a la hora de validar los resultados. De esta manera, el modelo a priori no tiene forma de saber cuáles son las secuencias del tipo *Hockey Stick*, lo que

también evita que se induzca un sesgo adicional por la forma en que se seleccionaron las secuencias durante el etiquetado manual. Dicho esto, se procede a explicar los pasos del método propuesto:

A. Preprocesamiento

Generar pares Cotización-*SKU* (concatenando las columnas) y asegurar que el conjunto de datos esté debidamente ordenado en los siguientes niveles:

Nivel 1: Pares Cotización-*SKU*

Nivel 2: Fecha de inicio (en los datos crudos los ítems no siempre están en orden)

B. Normalización de Duración y Cantidad

Para cada par Cotización-*SKU*:

Sumar las duraciones *D* de los ítems *i* y normalizarlas de modo que el total sea igual a 1.

$$D_{total} = \sum_{i} D_i$$

$$D_{normalizada_i} = \frac{D_i}{D_{total}}$$

$$\sum_{i} D_{normalizada_i} = 1$$

Calcular el centro temporal *C* de cada ítem *i*, con lo cual se generan los puntos del eje de abscisas *t*, que representa el transcurso del tiempo.

$$C_i = S_i + \frac{D_{normalizada_i}}{2}$$

Siendo S_i la suma acumulada de las duraciones normalizadas hasta el ítem *i*, sin incluirlo:

$$S_i = \sum_{j=1}^{i-1} D_{normalizada_j}, S_1 = 0$$

Registrar la cantidad *Q* del último ítem *n* de la secuencia y normalizarla a 1. Ajustar la cantidad de cada ítem en la misma proporción.

$$Q_{normalizada_i} = \frac{Q_i}{Q_n}$$

En general, la cantidad del último ítem Q_n será la cantidad máxima de cada secuencia, por lo que será igual a 1, pero se vio que en algunos casos la cantidad máxima se alcanza antes (lo cual ya es un comportamiento inesperado sobre el cual se profundizará más adelante). Es por esto que se introdujo el siguiente ajuste, el cual asegura que la cantidad máxima alcanzada en la secuencia se normalice a 1, independientemente de si esta se alcanza en el último ítem o no. Mediante este ajuste se terminan generando los puntos del eje de ordenadas *q*, que representa la cantidad suscripta en un momento *t* dado:

$$Q_{ajustada_i} = \frac{Q_{normalizada_i}}{\max\left(Q_{normalizada}\right)}$$

Véase una representación gráfica en la Figura 11. El ejemplo sería para un par Cotización-*SKU* con tres fases. Nótese que no todas las fases tienen la misma duración (t), pero su suma se normalizó a 1. En el caso de la tercera fase, el centro temporal es 0,8.

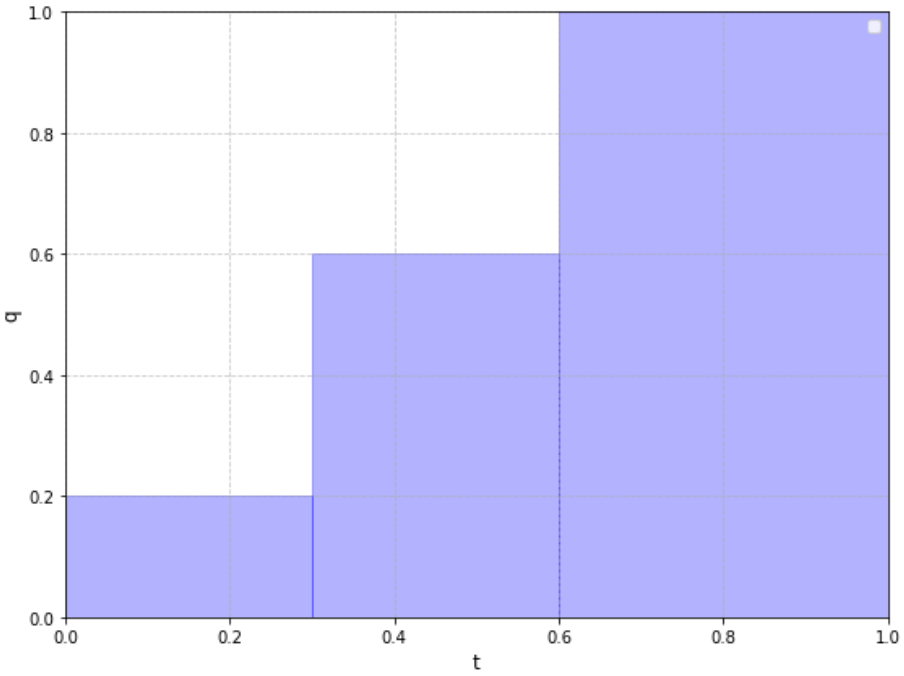

Figura 11. Representación gráfica de un par Cotización-*SKU* normalizado.

La figura 11 representa el caso típico en el sentido de que las cantidades (q) siempre van en aumento. Pero también podría llegar a darse el caso (inesperado) en el que otra fase sea la de mayor cantidad. De ser así el gráfico se vería como en la figura 12.

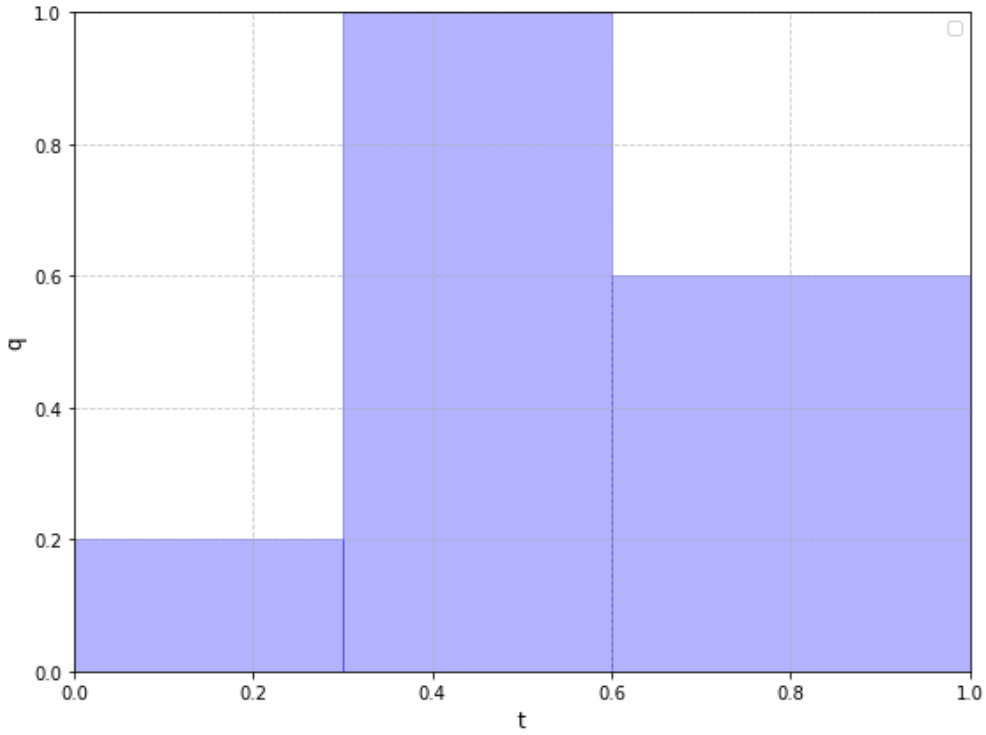

Figura 12. Representación gráfica de un par Cotización-*SKU* normalizado con comportamiento inesperado.

Este paso de normalización es absolutamente necesario, ya que permite la comparación de las secuencias que tienen largo variable en términos de cantidad de ítems y de duración real absoluta.

C. Entrenamiento del Modelo

Para cada *SKU*:

> Encontrar una función *f* que se ajuste lo mejor posible a todos los puntos obtenidos en el paso "B" en todas las cotizaciones del *SKU* en cuestión. Esta función representaría el comportamiento normal del *SKU*.

A modo de facilitar la comprensión de la idea, pasemos a verla sólo para un par Cotización-*SKU* en la figura 13, pero recordemos que el ajuste del modelo es a todas las ocurrencias (cotizaciones) del *SKU*.

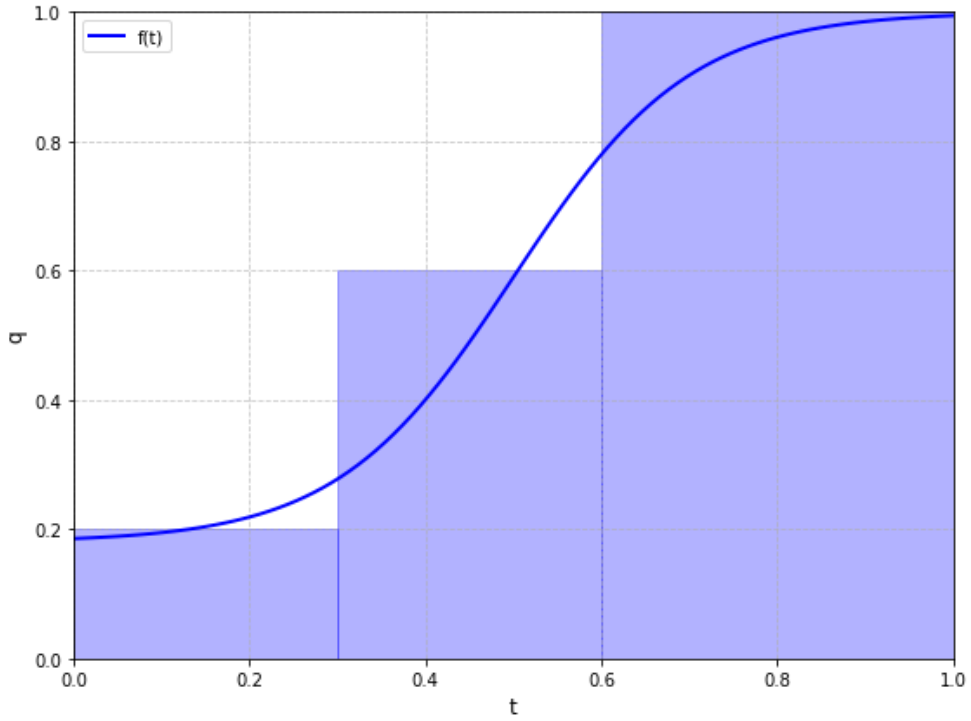

Figura 13. Representación gráfica de una función que busca ajustarse a un par Cotización-*SKU*.

D. Cálculo de Desvíos

Para cada par Cotización-*SKU*:

Comparar todos sus puntos obtenidos en el paso B contra la función obtenida en C para el *SKU* en cuestión y calcular un desvío D_i para cada uno de estos puntos. La figura 14 representa gráficamente esta idea para un único par Cotización-*SKU*. El desvío se muestra con las líneas punteadas rojas. Luego calcular el desvío promedio para todos los puntos.[13]

$$D_i = |Q_{ajustada_i} - f(C_i)|$$

$$D_{promedio} = \frac{\sum_{i=1}^{n} D_i}{n}$$

[13] En vez de utilizar el desvío promedio, alternativamente podría considerarse la utilización del desvío máximo de todos los puntos. Sin embargo, el desvío máximo podría sobreestimar el grado de atipicidad de secuencias normales, especialmente debido a que mayormente se encontraría en las primeras fases de corta duración de las suscripciones. Por el contrario, el desvío promedio brinda información más completa sobre el grado de atipicidad a lo largo de toda la duración de la suscripción.

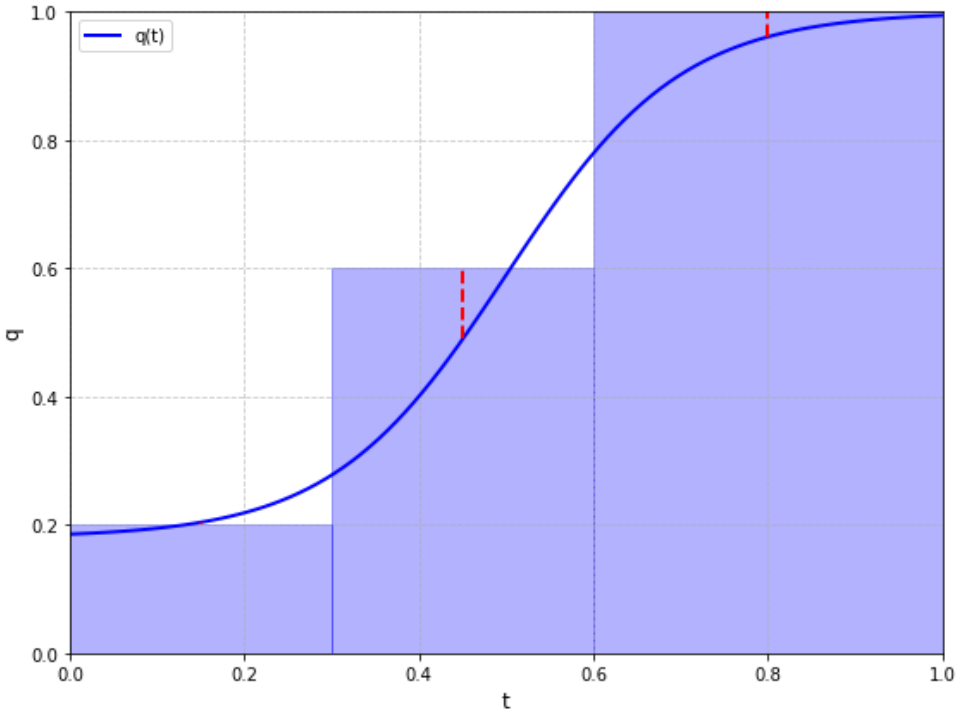

Figura 14. Visualización del cálculo de desvíos.

E. Compilado y Reordenamiento

Ordenar los pares Cotización-*SKU* de mayor a menor desvío promedio. De esta forma se obtiene una lista que muestra primero las secuencias que intuitivamente tienen mayor grado de atipicidad y luego las que tienen uno menor.

Adelantando algunos resultados, pero con la intención de facilitar la comprensión intuitiva de la solución, en la figura 15 se muestra una curva ajustada a todos los datos de un SKU, junto con los puntos de una secuencia normal en rojo, y los de una anómala en azul. Puede notarse que la distancia entre la curva y los puntos azules es mayor a la de los puntos rojos.

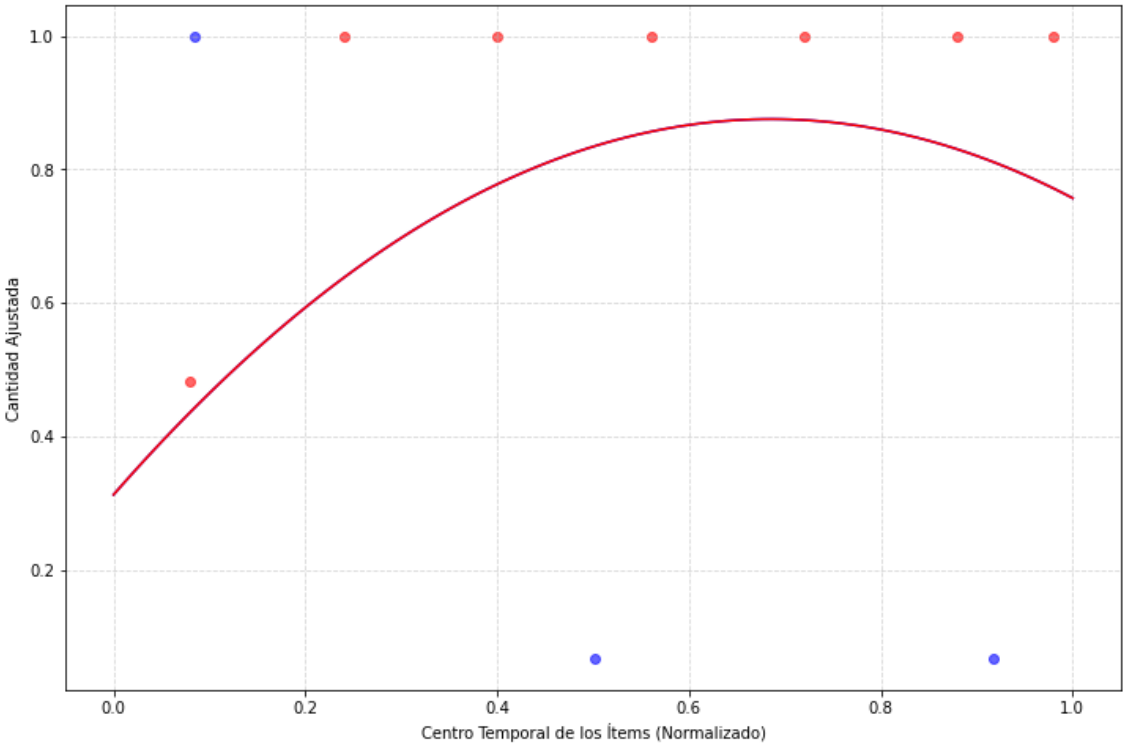

Figura 15. Comparación de una secuencia normal contra una anómala.

3.2.2 Detalles de Implementación

3.2.2.1 Exclusiones

Un detalle a tener en cuenta para la implementación es que las fases en la práctica se pueden generar en el sistema *CPQ* de dos formas: "Full" y "Delta". La diferencia se representa visualmente mediante dos gráficos en la figura 16.

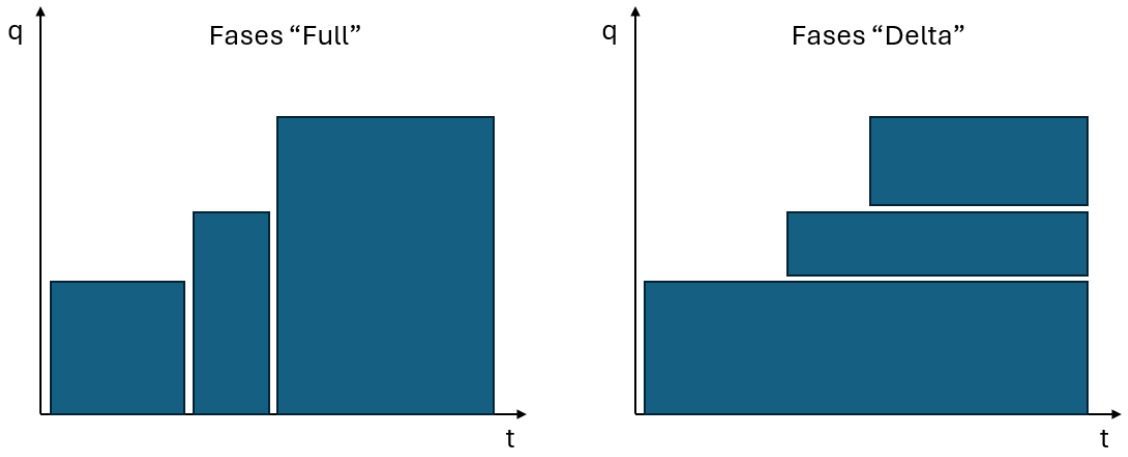

Figura 16. Comparación de estructuras con fases "Full" contra "Delta".

Puede observarse que las fases de ambos gráficos tienen la misma duración y mismas cantidades, pero en la estructura "Full" la cantidad de cada fase se obtiene con un único ítem representado por un rectángulo azul. Esa fase comienza en la fecha de inicio del ítem y termina en la fecha final del ítem. Por lo contrario, en la estructura "Delta", los ítems se van apilando para formar las fases, de modo que la cantidad de cada fase se obtiene a partir de todos los ítems activos en ese momento dado. Y la fecha final de todos los ítems es la fecha final del contrato.

Este último modo de estructurar contratos tiende a ser más complejo y difícil de interpretar a medida que un contrato tiene más fases, por lo que se está desincentivando su uso en la compañía. Debido a esta complejidad y a que se espera que el uso de este tipo de estructura vaya disminuyendo, se decidió excluir del análisis los pares Cotización-*SKU* que tengan esta estructura, lo cual se hizo identificándolos con el hecho de que tengan una misma fecha final en todos o la mayoría (más de 50% - también se vio que hay casos combinados delta-full) de sus ítems. Para poder analizarlos haría falta algún método de conversión a fases "Full", el cual no se desarrollará en este trabajo. Con esta exclusión el tamaño del conjunto de datos se redujo a 68.197 ítems correspondientes a 4.579 cotizaciones, lo que primero parece implicar una exclusión muy significativa considerando que inicialmente se contaba con 116.279 ítems correspondientes a 25.747 cotizaciones. Sin embargo, hay que tener en cuenta que con este paso también se eliminan todos los pares cotización-SKU que tienen un único ítem, es decir que no tienen fases. Al no tener fases, la cantidad es constante a lo largo de toda la duración del contrato, por lo que no hay anomalía alguna en esos casos y tiene sentido excluirlos. De los 48.082 ítems y 21.168 cotizaciones excluidas, 43.473 ítems y 20.741 cotizaciones tenían un único ítem. De este modo, el efecto neto de la exclusión únicamente por fases delta es de 4.609 ítems correspondientes a 427 cotizaciones, lo que termina siendo un impacto relativamente pequeño, aunque no despreciable.

3.2.2.2 Función de Ajuste a Utilizar y más Limpieza de Datos

Para el Paso C (Entrenamiento del Modelo) de la sección 3.2.1 es necesario elegir un tipo de función que se ajuste lo mejor posible a los datos. Para ello se probaron los tipos de funciones listados en la figura 17 y se midió su error de ajuste. Este error se calculó como la media de los $D_{promedio}$ que se habían definido en el paso D de la sección 3.2.1, de todos los pares cotización-*SKU.* De este modo, esta métrica nos indica qué tan grande es el error de ajuste a nivel agregado para todos los datos disponibles.

Función/Modelo	Ecuación	Error Medio Absoluto
Cúbica	$y = a * x^3 + b * x^2 + c * x + d$	0,1245
Cuadrática	$y = a * x^2 + b * x + c$	0,1263
Sigmoide	$y = \frac{L}{(1 + e^{-k * (x - x_0)})}$	0,1298
Lineal	$y = a * x + b$	0,1328
Exponencial[14]	$ln(y) = a * x + b$	0,1739

Figura 17. Error Medio Absoluto de las funciones probadas.

Cabe destacar que se optó por utilizar el Error Medio Absoluto como métrica por sobre otras posibles como errores relativos/porcentuales o la Raíz del Error Medio Cuadrático por los siguientes motivos:

1. Las duraciones y cantidades están normalizadas de antemano
2. Fácil interpretabilidad en las unidades normalizadas
3. Robustez del modelo (evitar eventuales errores al dividir por cero)

Volviendo a la elección de la función a utilizar, lo ideal sería poder elegir el tipo que más se adapte al comportamiento normal de cada *SKU*, pero por motivos de simplicidad, se decidió elegir un único tipo para todos los *SKUs*. Podríamos utilizar el Error Medio Absoluto de la figura 17 como único criterio para esta elección, pero también sería prudente considerar la forma que cada función tiene, comparándola con el comportamiento normal esperado por el negocio, y también teniendo en cuenta que se debe evitar el sobreajuste, especialmente en casos con pocos datos. Es por esto que para mejorar la comprensión de cómo se estaba comportando cada tipo de función se generaron mapas de densidad con las funciones ajustadas de todos los *SKUs*. En las figuras 18, 19 y 20 podemos ver esto para las funciones Cuadrática, Cúbica y Sigmoide (las tres con el menor error), respectivamente.

[14] Nótese la transformación de la variable dependiente. Ajustando una función lineal al logaritmo natural de la variable dependiente, el modelo capta una relación exponencial en la escala original.

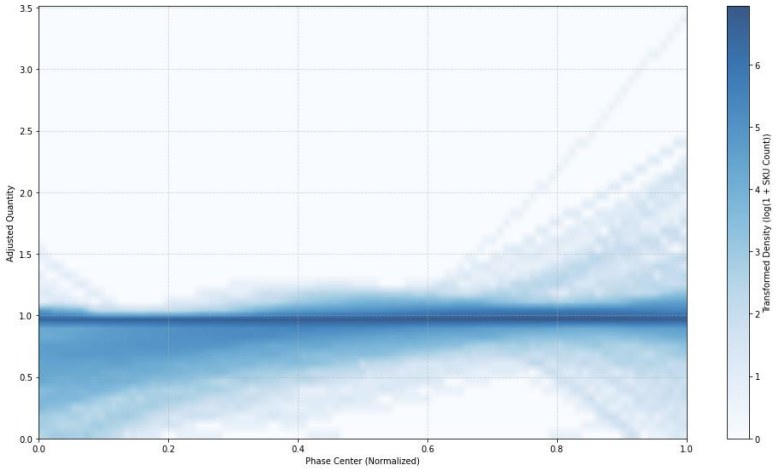

Figura 18. Mapa de Densidad de Funciones Cuadráticas.

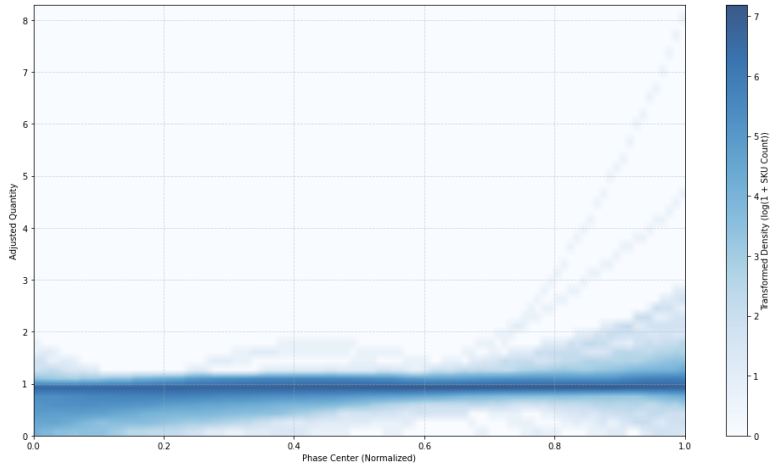

Figura 19. Mapa de Densidad de Funciones Cúbicas.

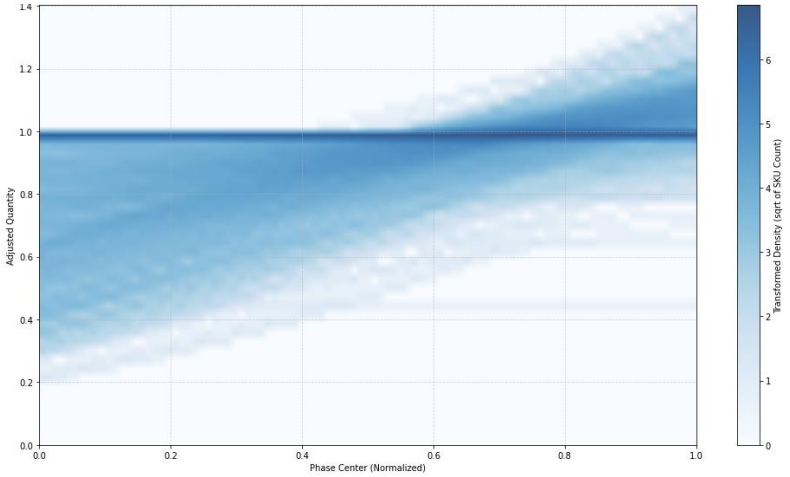

Figura 20. Mapa de Densidad de Funciones Sigmoides.

Más allá de las distintas formas de los diferentes tipos de funciones, lo que todas tienen en común es una fuerte concentración de datos en torno a la cantidad normalizada 1 a lo largo de toda la duración del contrato, es decir, lo más común es que en un contrato la cantidad se mantenga constante a lo largo de toda su duración. Esto explica también por qué la función lineal tenía un error de ajuste tan bajo, que sorprendía a priori. A pesar de que ya se habían excluido las secuencias con un único ítem (que consecuentemente también tenían cantidad constante), estas secuencias con cantidad constante tienen varios ítems, todos con la misma cantidad.

Como este comportamiento también es totalmente normal y no conlleva ningún riesgo para el negocio, los datos correspondientes a estos casos se eliminaron para obtener un ajuste que sea más útil para nuestro objetivo. Dicho de otra forma, solamente se dejaron los datos de contratos con variaciones de cantidad.

Luego de eliminar estos casos, el tamaño del conjunto de datos relevante se reduce aún más, quedando 23.714 ítems, correspondientes a 1.629 cotizaciones. Con este paso, también se eliminan los pares Cotización-*SKU* que solo tienen un ítem, los cuales se habían identificado como irrelevantes, por lo menos para la anomalía del tipo *Hockey Stick*. Con este conjunto de datos reducido Los mapas de densidad se ven así (Figuras 21, 22 y 23):

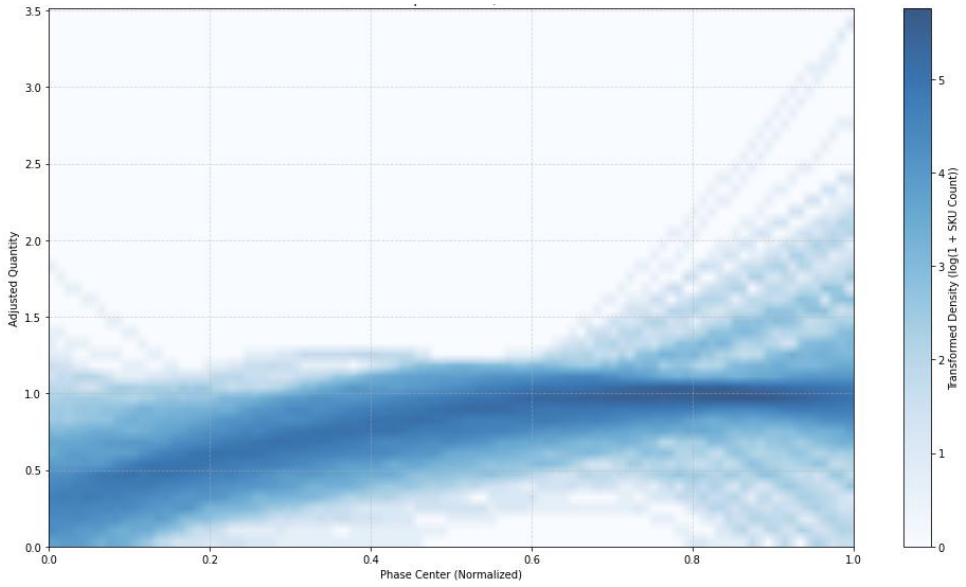

Figura 21. Mapa de Densidad de Funciones Cuadráticas.

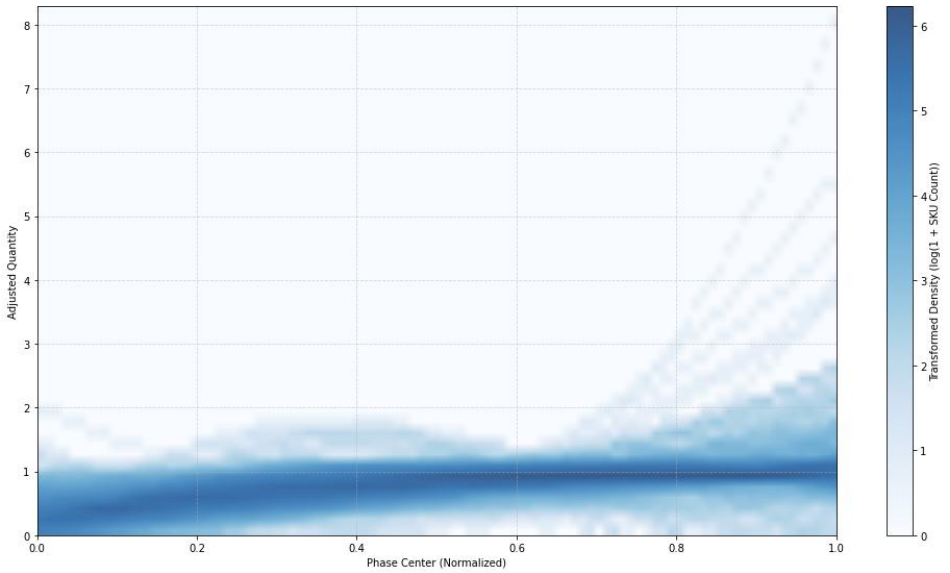

Figura 22. Mapa de Densidad de Funciones Cúbicas.

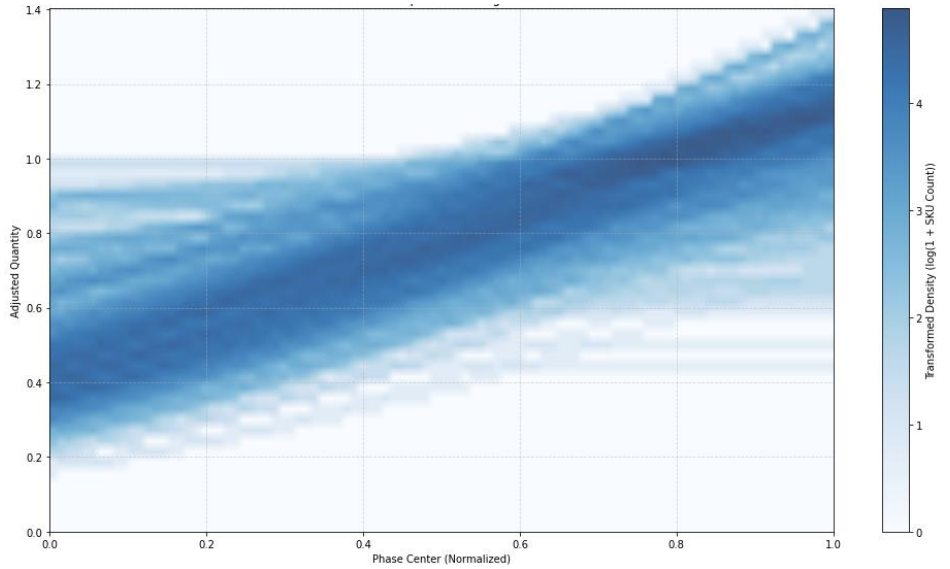

Figura 23. Mapa de Densidad de Funciones Sigmoides.

Si bien ahora puede notarse la convergencia hacia la cantidad normalizada 1, ésta ya no es constante a lo largo de toda la duración de los contratos. También puede verse que la tendencia general es que las cantidades suscriptas crezcan a lo largo de los contratos, lo que también es el comportamiento esperado por el negocio. Con los nuevos datos, se procede a recalcular el error de ajuste a nivel agregado en la Figura 24 y se agrega el Desvío Estándar, que proporciona además información acerca de qué tanto los errores de ajuste difieren entre distintos pares cotización-*SKU*:

Función	Error Medio Absoluto	Desvío Estándar
Cúbica	0,1682	0,1561
Cuadrática	0,1725	0,1516
Sigmoide	0,1975	0,1548

Figura 24. Error Medio Absoluto y Desvío Estándar de las funciones seleccionadas.

En general podemos ver que el error medio es mayor que antes debido a que eliminando los casos constantes es más difícil ajustarse a los datos que quedan. La Función Cúbica es la que menor error medio tiene, pero mayor desvío. Esto se debe a que al tener un mayor grado, puede ajustarse mejor a los datos, pero se corre mayor riesgo de sobreajustar. Gráficamente, en la figura 22 llama la atención como la función para algunos *SKUs* se escapa a valores muy superiores a 1 hacia el final de la duración.

La función sigmoide es la que tiene el mayor error, y un desvío muy cercano al de la función cúbica. Gráficamente (Figura 23) podemos ver que el modelado es mucho más simple que el de la cúbica. En el caso de la cuadrática el error se encuentra entre ambas otras, pero es apenas superior al de la cúbica. El desvío en cambio es el menor por un amplio margen relativo, indicando mayor consistencia en las predicciones. Es por este motivo que se decide continuar el análisis con un modelo cuadrático para todos los *SKUs*. En el análisis gráfico de la figura 21, podemos ver que hay curvas que se salen de la zona con mayor densidad, pero luego de un análisis más detallado, se concluye en que estas curvas corresponden a pares Cotización-*SKU* con pocos puntos de entrenamiento, por lo que las funciones se sobreajustan a ellos.

Un punto a aclarar, naturalmente hay valores atípicos, numéricamente distantes del resto de los datos procesados, y estos son los que se desean descubrir mediante el presente trabajo. En el espacio normalizado de tiempo y cantidad un valor atípico va a venir dado por su contexto, es decir, ¿qué cantidades había antes y después? ¿Por cuánto tiempo se tuvo esa cantidad extraordinaria? Como se presume que estos casos son infrecuentes y no todos están identificados a priori, el entrenamiento del modelo se hace sin excluir valores atípicos. Lo que sí debe tenerse en cuenta es la sensibilidad de las funciones ajustadas a los valores atípicos. Particularmente en nuestro caso, la función cuadrática parece ser que mejor balancea la sensibilidad ante valores atípicos con la precisión del ajuste. Esto se debe a que tiene un grado polinómico menor que el de la función cúbica, que es más flexible y propensa a sobreajustes y distorsiones, mientras que la función sigmoide tiene un rango de salida más limitado, lo que puede restringir su capacidad de modelar comportamientos particulares de cada *SKU.*

3.2.2.3 Formalización

Ahora que ya se seleccionó una función a utilizar, se procede a formalizar el problema:

Sea para un par Cotización-*SKU* dado *j*:

- n_j la cantidad de ítems del par Cotización-*SKU j.*
- $x_{i,j}$ el centro temporal para el *i*-ésimo ítem del conjunto normalizado de valores entre 0 y 1 del par Cotización-*SKU j*.[15]
- $Q_{i,j}$ la cantidad normalizada y ajustada para el *i*-ésimo punto.
- $f(x_{i,j}) = ax_{i,j}^2 + bx_{i,j} + c_j$ la cantidad predicha para el *i*-ésimo punto de datos en base a la función con los parámetros *a, b* y *c* que se estimó para el *SKU* en cuestión con los datos de todas las cotizaciones en las que aparece.
- $D_{i,j} = |Q_{i,j} - f(x_{i,j})|$ el desvío absoluto para el *i*-ésimo punto respecto a su comportamiento normal/esperado, representado por $f(x_{i,j})$.
- $D_{promedio,j} = \frac{1}{n_j} \sum_{i=1}^{n_j} (D_{i,j})$ el desvío absoluto promedio del par Cotización-*SKU j*.

En la figura 25 se puede ver un ejemplo con todos los puntos disponibles para un *SKU*, donde los correspondientes a cada cotización se muestran con un color distinto. Es decir, los puntos de un mismo par Cotización-*SKU* se muestran del mismo color. Los desvíos $D_{i,j}$ se calculan entre los puntos y la curva roja que representa el comportamiento normal esperado de cada *SKU*.

[15] Aclaración: En la sección 3.2.1 esta variable se había identificado con la letra *C*. Aquí se optó por remplazarla con la letra *X* para distinguirla con mayor facilidad del parámetro *c* de la función cuadrática.

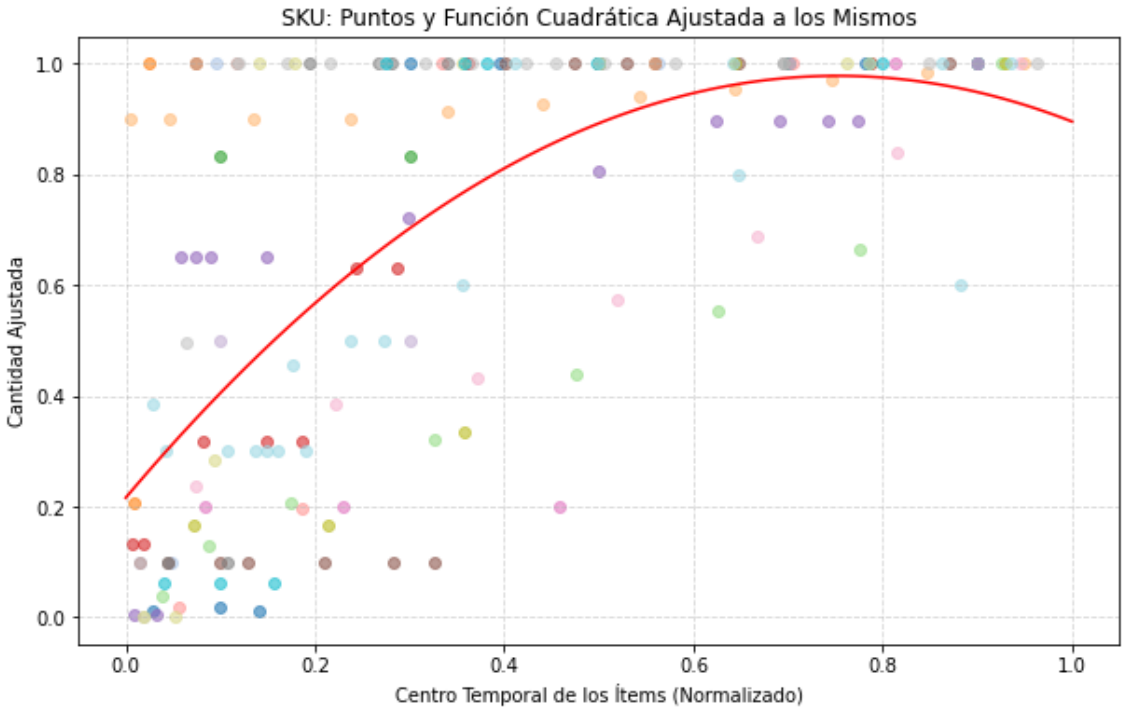

Figura 25. Ejemplo de función cuadrática ajustada a los puntos de un *SKU*.

Una vez que se computó el desvío para todos los puntos de cada *SKU*, se calcula el error o desvío medio $D_{medio,j}$ para cada par Cotización-*SKU*. Ordenando éstos de mayor a menor podemos encontrar los casos que a priori serían anómalos en los primeros lugares de la lista.

3.2.3 Limitaciones del Modelo

Con la solución propuesta en la sección anterior se obtuvo un desvío promedio para cada secuencia disponible en el conjunto de datos remanente. Si bien una verificación preliminar de estos resultados ya permite ver que las secuencias con mayores desvíos promedio tienden a tener comportamientos que efectivamente llaman la atención (se profundizará más en la sección 4, Resultados), en esta verificación también se vio que hay secuencias con desvíos promedio altos, que tienen comportamientos totalmente normales ante los ojos de un analista.

Ahora surgen preguntas como las siguientes: ¿A partir de qué valor de desvío una secuencia se debe considerar como anómala? ¿Todas las secuencias por encima de cierto umbral realmente son anómalas? ¿Todas las anomalías son similares? ¿Qué tipos de anomalías se encuentran? ¿Cuáles pueden tener un impacto negativo en el negocio? Estas son preguntas que esta primera etapa de regresión no puede responder, por lo que el análisis se continúa en la siguiente etapa.

3.3 Segunda Etapa: Segmentación

3.3.1 Problemática y Solución Propuesta

Esta segunda etapa consistirá esencialmente en comprender mejor los resultados de la etapa anterior para poder distinguir con mayor certeza las secuencias anómalas. En primer lugar, veamos cómo se distribuyen los desvíos promedio obtenidos (Figura 26).

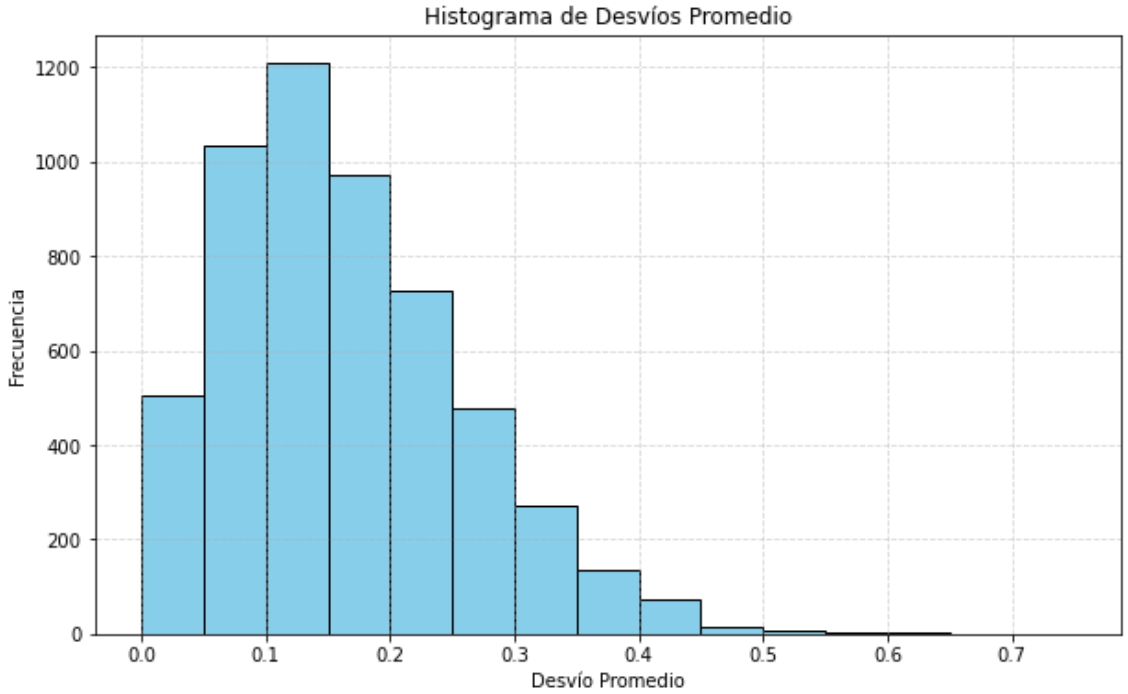

Figura 26. Distribución de desvíos medios de todas las secuencias analizadas.

Puede verse que la mayoría de las secuencias (alrededor de 1.200) tienen un desvío medio de entre 0,05 y 0,2 mientras que cruzando sobre 0,3 las frecuencias se hacen notablemente menores, indicando comportamientos menos frecuentes. La secuencia más extrema tiene un desvío de 0,69. Ahora, ¿qué pasa cuando sólo se observan los casos que se habían etiquetado como *Hockey Stick*?

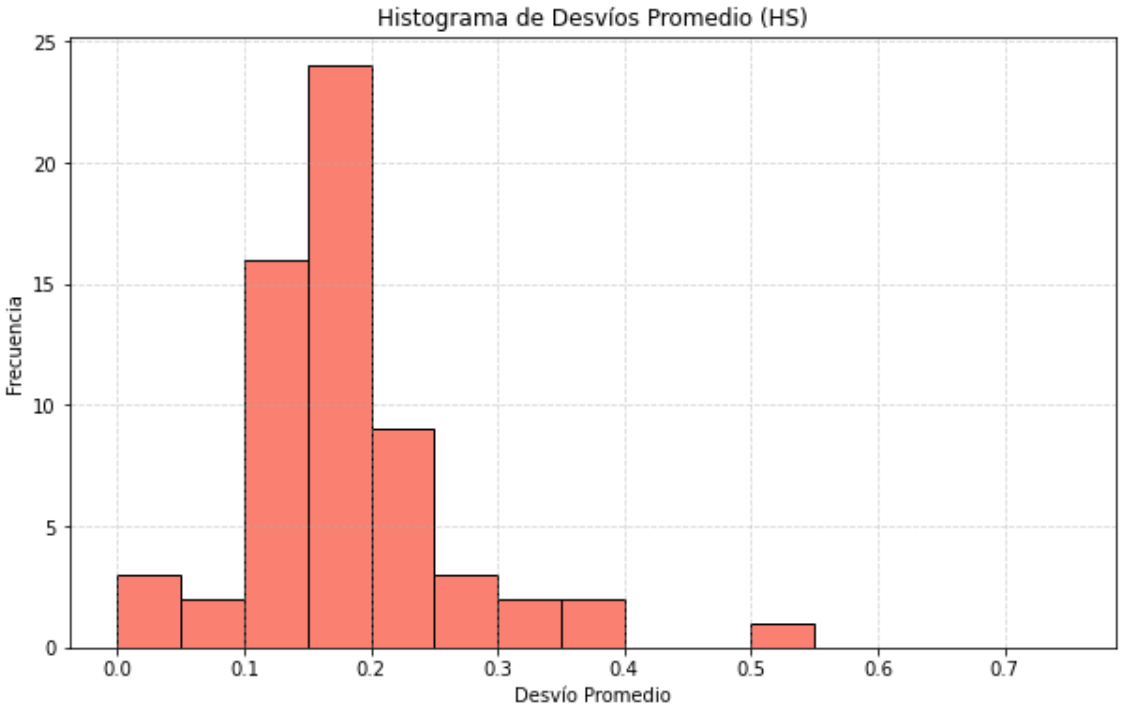

Figura 27. Distribución de desvíos medios de las secuencias etiquetadas como *Hockey Stick*.

En la figura 27 puede notarse que el centro de la distribución se movió a valores levemente más altos, pero en incluso en los intervalos más bajos del histograma sigue habiendo casos que fueron etiquetados como riesgosos. Esto implica que sería poco prudente trabajar con un umbral de desvío como único indicador para detectar casos anómalos.

Por este motivo se decidió implementar una etapa de agrupamiento no supervisado, para no solamente detectar las secuencias anómalas, sino que también poder distinguir distintos tipos de anomalías. Para ello se utilizará el algoritmo de k-medias, debido a su simplicidad, facilidad de implementación, escalabilidad y flexibilidad con el número de clústeres. También se explorarán dos variantes que se explicarán a continuación.

Para llevar a cabo la segmentación de las secuencias, el algoritmo tomará como inputs las siguientes variables. Algunas de ellas ya estaban en el conjunto de datos, mientras que otras fueron creadas para este propósito:

(1) Cantidad de ítems
(2) Cantidad de niveles de volumen distintos
(3) Rango entre el nivel de volumen inferior y el superior
(4) Duración total de la secuencia (días)

(5) Desvío estándar de las duraciones de los ítems (días)
(6) Duración del último ítem (días)
(7) Duración porcentual de (6) respecto a la duración total de la secuencia
(8) Duración de los todos los ítems en el nivel de volumen máximo
(9) Duración porcentual de (8) respecto a la duración total de la secuencia
(10) Duración de los todos los ítems en el nivel de volumen mínimo
(11) Duración porcentual de (10) respecto a la duración total de la secuencia
(12) Término "a" de la función cuadrática ajustada al *SKU*
(13) Término "b" de la función cuadrática ajustada al *SKU*
(14) Término "c" de la función cuadrática ajustada al *SKU*
(15) Desvío medio obtenido en la etapa de regresión
(16) Valor neto total ($)

Se espera que estas variables le puedan dar más información al algoritmo para poder segregar los casos con mayor facilidad. Estas variables también son escaladas para facilitar su procesamiento, ya que tienen unidades muy distintas (desde valores monetarios expresados en millones, hasta desvíos expresados en decimales). Luego es necesario determinar la cantidad *k* de clústeres o grupos, balanceando tener la granularidad necesaria para distinguir las anomalías, contra tener una cantidad lo más baja posible de clústeres para facilitar el análisis lo más posible.

Vale la pena aclarar que si bien los métodos basados en distancias suelen perder robustez en contextos de alta dimensionalidad (16 variables en este caso), en la primera variante que se expondrá, se trabajará con todas las variables para tener una referencia base. En la segunda variante se terminará trabajando con técnicas para reducir la dimensionalidad.

3.3.1.1 Primera Variante: K-Medias Estándar

3.3.1.1.1 Detalles de implementación

En esta primera aplicación para determinar *k* se utilizó el *Método del Codo*, en el cual se calcula la llamada Inercia para distintos valores de *k.* La inercia representa una medida de qué tan bien los datos se agrupan alrededor de los centroides de cada clúster y se define como:

$$Inercia = \sum_{i=1}^{n} \min_{\mu_j} || \ x_i - \mu_j \ ||^2$$

Donde:

n es la cantidad de secuencias

x_i es la secuencia *i*-ésima

μ_j es el centroide del clúster *j*-ésimo

$\min_{\mu_j} || x_i - \mu_j ||^2$ es la distancia Euclídea cuadrada entre una secuencia y su centroide más cercano.

De modo que valores bajos indican que los datos están más cerca de sus respectivos centroides, y se ajustan mejor. Si bien este método se presta a cierta ambigüedad y no considera qué tan bien están separados los clústeres, es un método simple y fácil de entender intuitivamente. A continuación, se grafica la inercia para distintos valores de *k.*

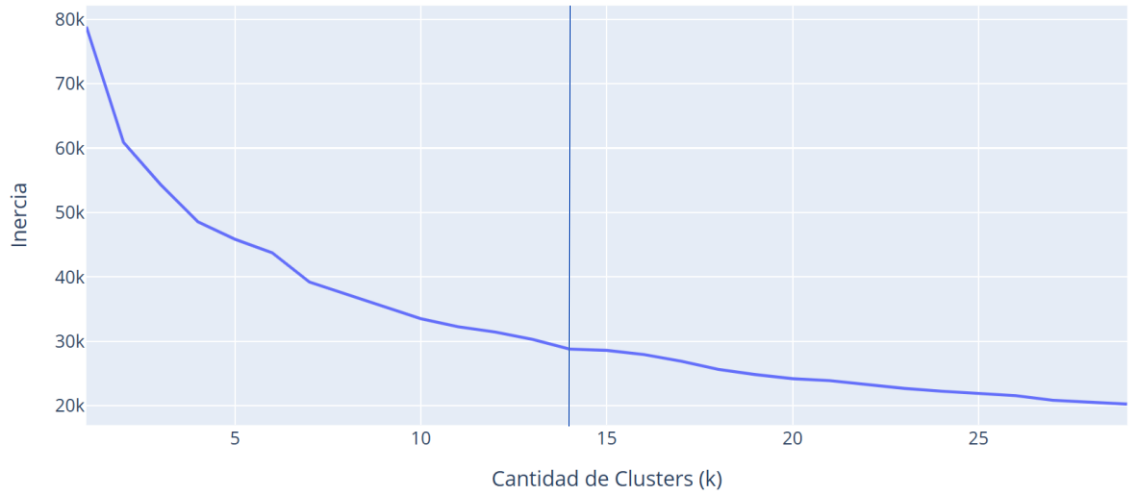

Figura 28. Determinación visual de *k* con el *Método del Codo*

Inspeccionando la Figura 28 visualmente y buscando codos a partir de los cuales un aumento de *k* no genera una reducción significativa de la Inercia, se encontraron tres candidatos: k = 7, 10 ó 14. Para decidir, se realizó una inspección preliminar de la asignación de las secuencias a los clústeres y se vio que tanto *k* = 7 y 10 no proveían el grado de separación detallado que sí proveía *k* = 14. Luego lo veremos en más detalle, pero por ejemplo, con *k* = 14 todos los casos *Hockey Stick* quedaban agrupados en un único clúster, mientras que con *k* = 7 o 10 los casos se repartían en más de un clúster. Por este motivo se decidió que *k* = 14 es el valor con el cual se trabajará en esta variante. Es decir, las secuencias ordenadas de mayor a menor desvío promedio se repartirán en catorce grupos, esperando que cada grupo represente secuencias con características particulares, e idealmente algunos de esos grupos con un único tipo de anomalía.

3.3.1.1.2 Análisis de Resultados

Combinando la etapa de regresión con esta etapa de agrupamiento no supervisado se obtuvo una lista de secuencias ordenada de mayor a menor desvío promedio y agrupada en 14 clústeres (del 0 al 13). Estos datos se salida se analizaron manualmente prestando especial atención a las secuencias con mayor desvío de cada clúster para comprender su naturaleza con el menor esfuerzo posible. De no tener las secuencias ordenadas, sería mucho más difícil identificar anomalías, ya que algunos clústeres resultaron tener anomalías pero sólo en algunas de sus secuencias, y como se esperaba, estas tendían a ser las que tenían el mayor desvío promedio.

Para acotar la lista y facilitar el análisis de los resultados, en esta etapa se decidió observar únicamente secuencias de contratos correspondientes a contratos firmados en 2024. Desde una perspectiva de negocio esto tiene el propósito de evaluar comportamientos que hayan sucedido hace relativamente poco y que por lo tanto, tienen mayor probabilidad de estar sucediendo actualmente. Es decir, el modelo se entrenó con el conjunto de datos completo, pero a la hora de evaluar resultados, solo observaremos los de 2024. A continuación, vamos a pasar a ver esto más detalladamente, clúster por clúster. En cada uno se comenzó revisando las secuencias con mayor desvío, las cuales efectivamente tendían a tener más anomalías, y a medida que se iba llegando a casos con menores desvíos, era cada vez más difícil encontrar anomalías reales o que puedan llegar a tener un impacto negativo en el negocio, si bien sí se encontraron algunas. Cabe aclarar que entre los valores más altos de desvíos también hay casos que no presentan anomalías reales, lo que se puede confirmar mediante la revisión de un analista. Por ejemplo, en la figura 29 se muestra una secuencia con dos ítems, que es una de las que mayor desvío promedio tiene (0,43 – verifíquese en la figura 26 que es en la parte superior de la distribución), a pesar de que revisándola en detalle no hay nada que llame la atención. La primera fase, probablemente de implementación, toma el primer 20% (centro en 0,1) de la duración total con poca cantidad, mientras que el último 80% (centro en 0,6) de la duración es en la cantidad máxima suscripta. Recordemos que el comportamiento esperado por el negocio es que las suscripciones tengan cantidades crecientes (o constantes – aunque estos casos ya fueron eliminados del conjunto de datos). Esto lo vemos reflejado en la curva cuadrática ajustada a los datos de todas las secuencias del *SKU* en cuestión.

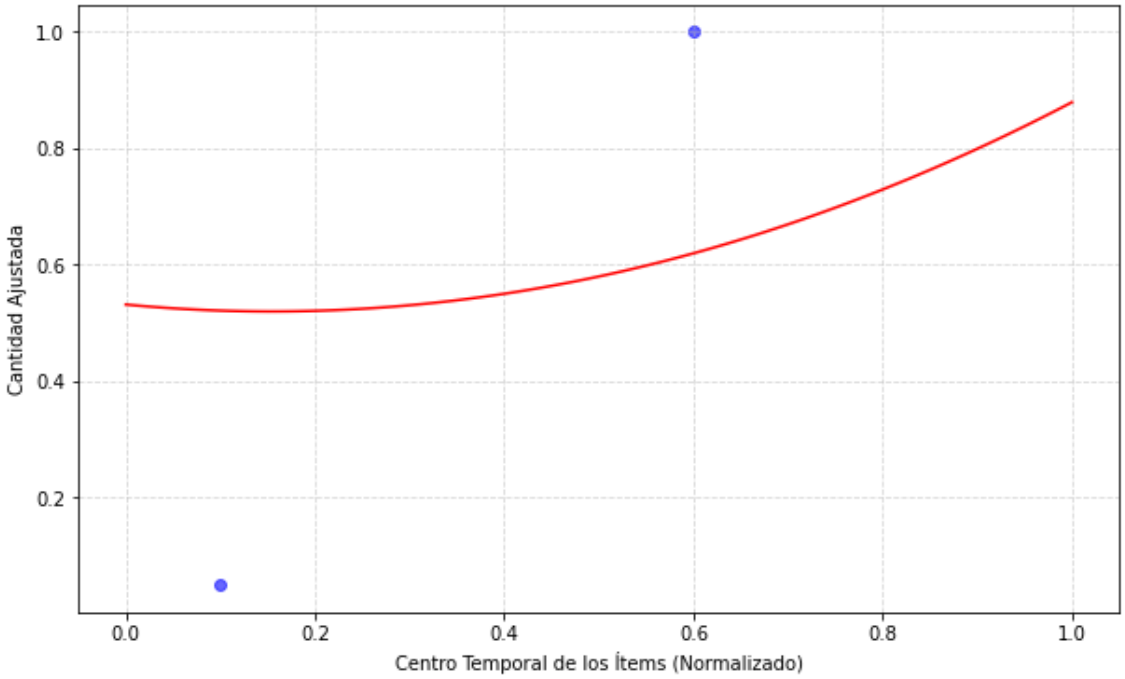

Figura 29. Secuencia con alto desvío calculado, pero normal para el negocio.

Clúster 0

Es el que contiene las secuencias con mayores desvíos. En el top 10 de los desvíos de todos los clústeres, cuatro corresponden a este. Una característica común de este clúster es que sólo contiene secuencias que se mantienen dentro de un único nivel de volumen. Es decir que puede haber variaciones en la cantidad contratada a lo largo del contrato, pero estas variaciones se mantienen siempre dentro del mismo nivel de volumen, probablemente porque son *SKUs* con un único nivel disponible. Pasemos a ver algunos ejemplos de anomalías encontradas (más detalles disponibles en el apéndice):

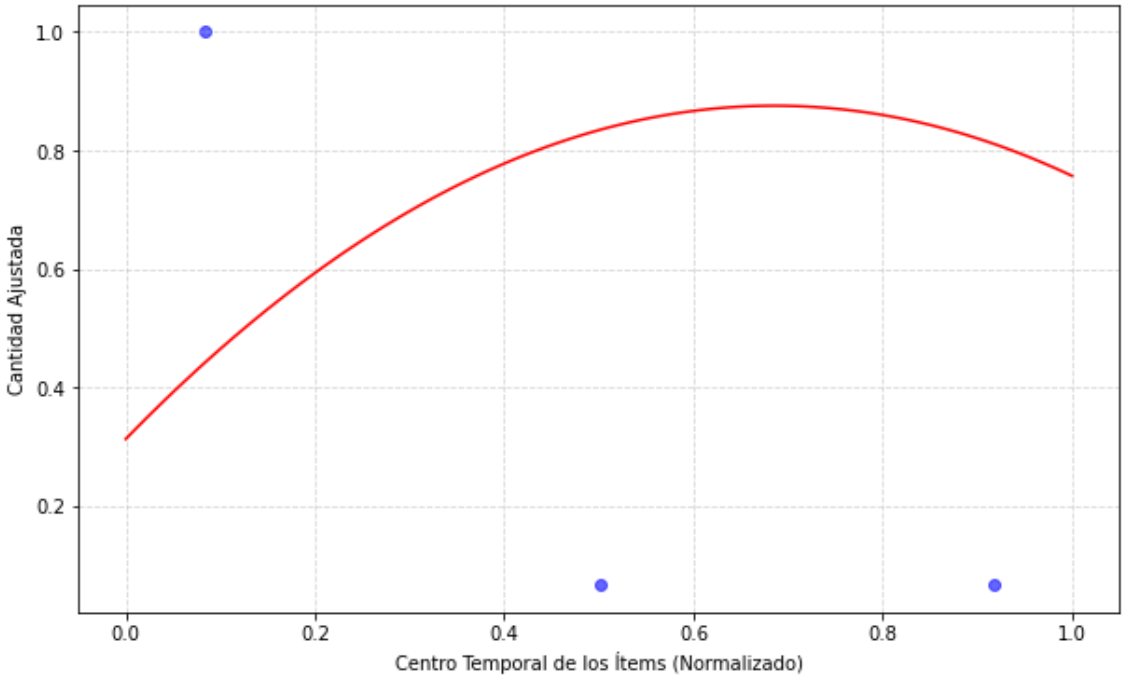

Figura 30. Secuencia con cantidad decreciente al principio del contrato.

La figura 30 muestra la secuencia que tiene el mayor desvío promedio de todos los clústeres (0,69). Puede verse cómo la cantidad máxima se alcanza en el primer ítem, lo que ya va en contra del comportamiento esperado por el negocio, en el cual las cantidades crecen. Revisando los datos originales esta cantidad alcanzada en la primera fase corresponde a 15 usuarios con una duración de 184 días, mientras que la segunda fase comprendida por los dos ítems siguientes tiene 911 días con un único usuario. Es decir, con este caso ya se encontró un tipo de anomalía nueva, caracterizada por una alta cantidad suscripta durante un período de tiempo corto al principio del contrato, respecto a la duración restante. En el clúster se encontraron siete casos más con el mismo tipo de comportamiento, aunque con un desvío promedio menor. Sin embargo, no se descarta la presencia de más ocurrencias, ya que la revisión manual no fue exhaustiva, es decir no se llegaron a revisar en detalle las secuencias con desvíos más pequeños.

En la figura 31 se puede ver un comportamiento similar en el sentido de que la cantidad suscripta va decreciendo. La diferencia aquí es que la caída se da hacia el final de la suscripción.

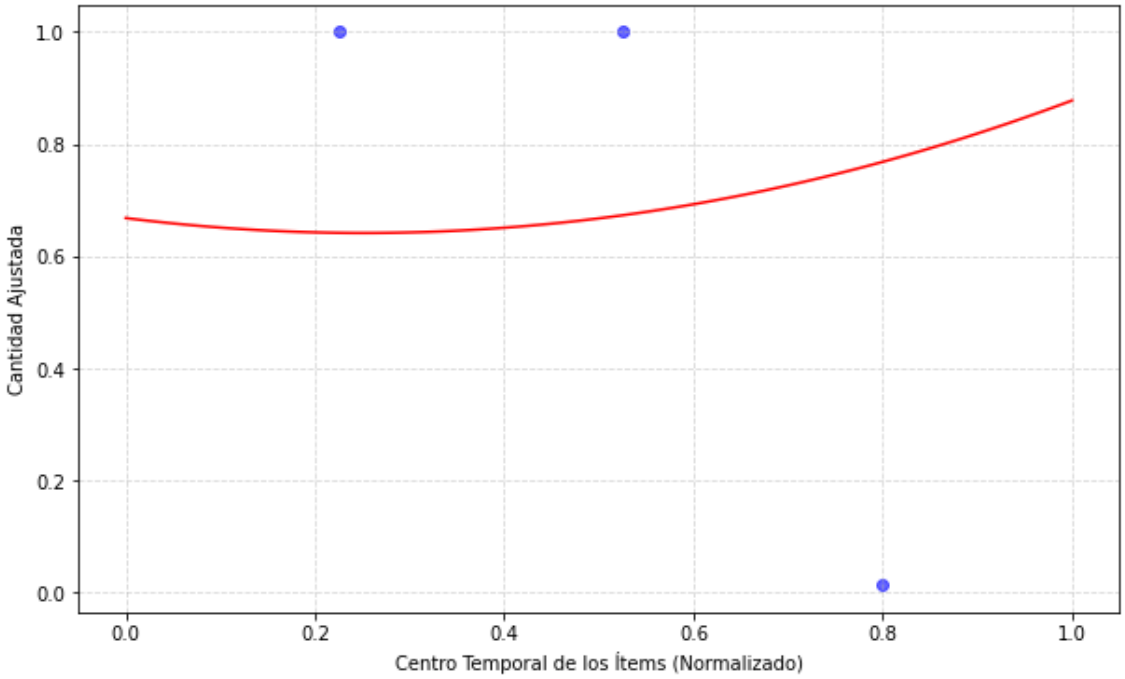

Figura 31. Secuencia con cantidad decreciente hacia el final del contrato.

En este caso la cantidad suscripta pasa de 350 unidades durante el primer 60% de la duración a 5 unidades en el último 40%. Con este tipo de comportamiento se encontraron 12 casos en este clúster, nuevamente sin descartar la posibilidad de que haya más ocurrencias no vistas.

El caso inverso lo vemos en la figura 32, que muestra una secuencia de 1.816 días de los cuales los últimos 242 días tienen 51 unidades suscriptas, mientras que el resto de la duración tiene 3 unidades. Esta es la secuencia que tiene el segundo mayor desvío de todos los clústeres (0,62). En este clúster sólo se encontró una secuencia más con un comportamiento similar. Nótese que este comportamiento es similar al *Hockey Stick* que se describió en la sección 2.1, particularmente en la figura 5. Pero a diferencia de este, la variación de cantidad ocurre dentro del mismo nivel de volumen, por lo que en principio no hay riesgo de pérdidas por costos no cubiertos. Sin embargo, sigue llamando la atención el motivo de este comportamiento.

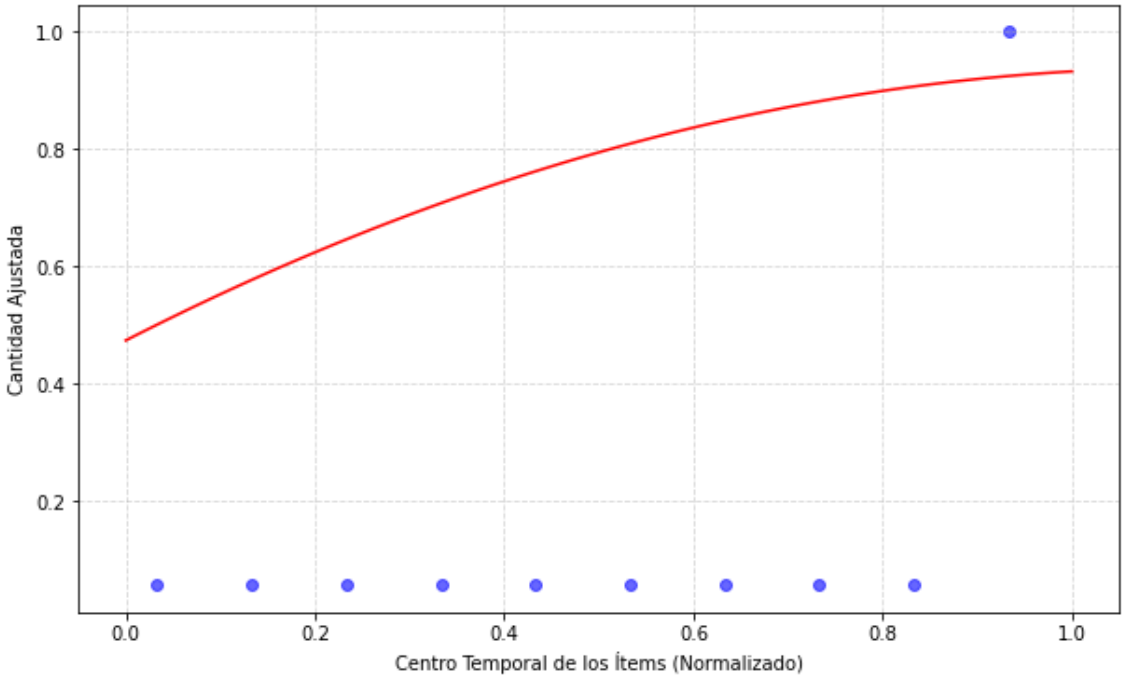

Figura 32. Secuencia con fuerte incremento de cantidad hacia el final del contrato.

Otro tipo de anomalía encontrada es la que se muestra en el ejemplo de la Figura 33, donde la cantidad aumenta, como es esperado, pero luego vuelve a caer significativamente. De este tipo, se encontraron cinco ocurrencias.

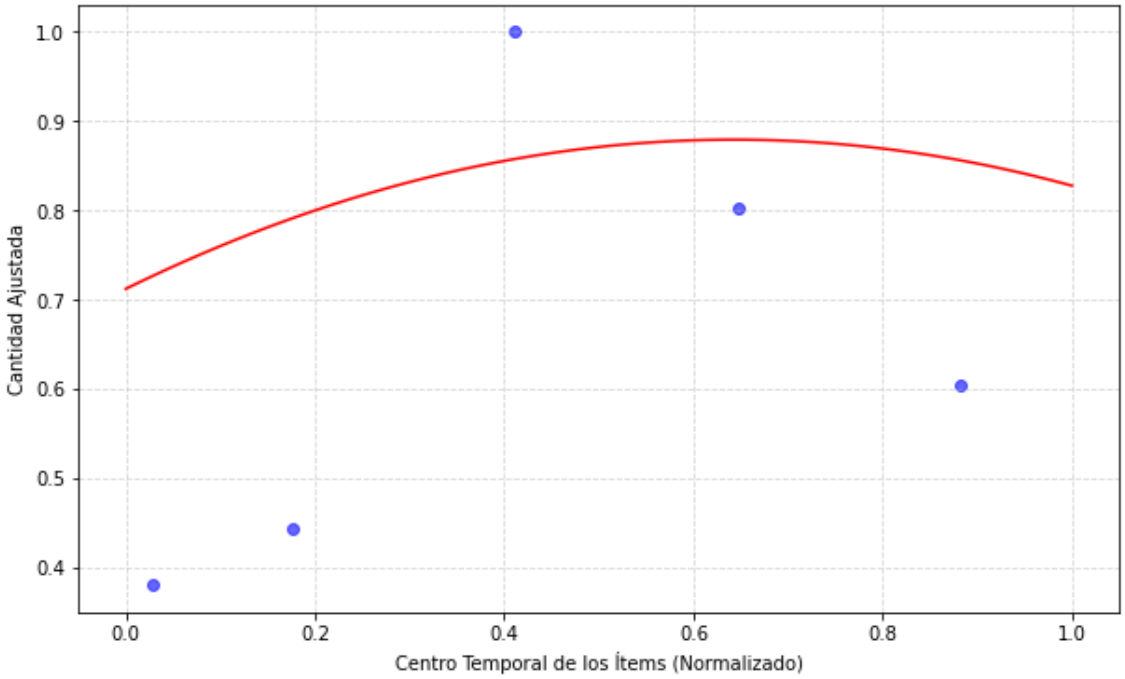

Figura 33. Secuencia con cantidad creciente al principio y decreciente hacia el final.

Y el último tipo de anomalía que se encontró en este clúster es el de cantidad decreciente a lo largo de todo el contrato, cuyo ejemplo se muestra en la Figura 34.

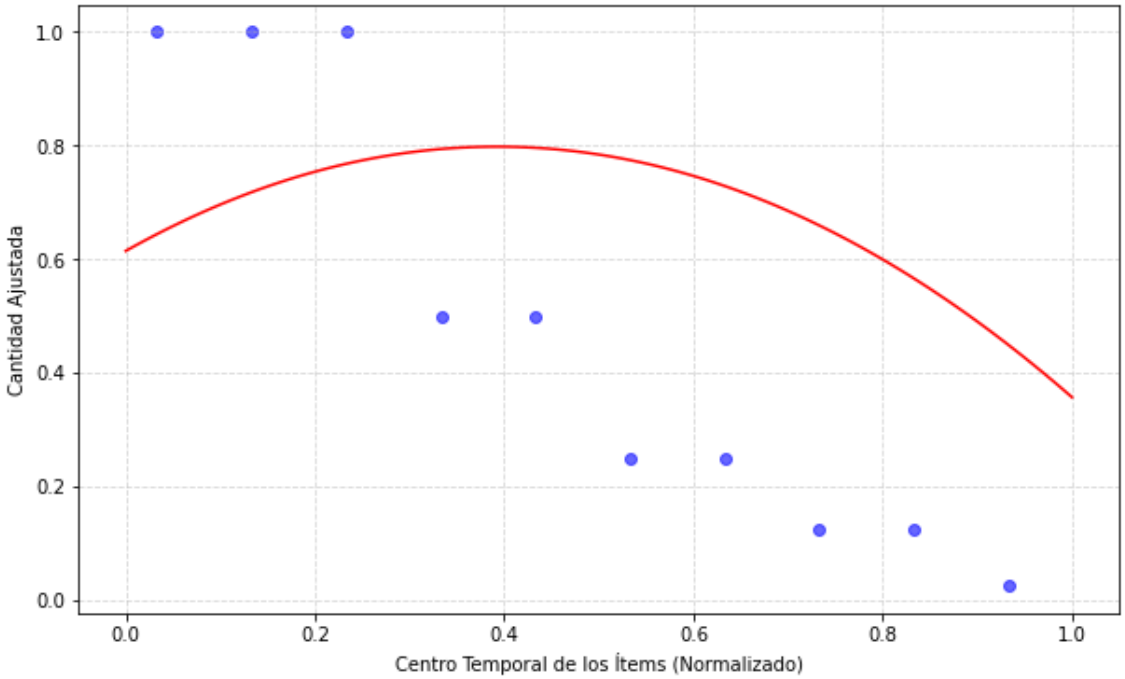

Figura 34. Secuencia con cantidad decreciente a lo largo de toda su duración.

En este caso también llama la atención que la curva que se había ajustado a todas las secuencias de este *SKU* también es decreciente, de modo que en este *SKU* ese también pareciera ser el comportamiento más común. Queda para investigar si ese es el verdadero comportamiento esperado por el negocio.

En resumen, este Clúster 0 contiene secuencias con varios comportamientos que llaman la atención, aunque por lo menos, como siempre se comercializan en un mismo nivel de volumen, no corren riesgo de generar pérdidas, considerando lo explicado de los casos *Hockey Stick* en la sección 2.1. Sí surgen preguntas respecto a por qué aparecen estos comportamientos. ¿Por qué existen los saltos vistos? ¿Se vende de más en ciertos períodos? ¿Qué pasa con las renovaciones de los contratos? ¿Hay motivos legítimos para estos comportamientos? ¿Hay posibilidad de que mayor proporción de la duración de los contratos esté más cerca de la cantidad máxima alcanzada aumentando la facturación? Éstas son preguntas que se podrían abordar en una siguiente etapa.

Clústeres 1, 2, 4, 5, 8, 12 y 13

En la revisión de estos clústeres no se encontraron anomalías a pesar de que algunas pocas secuencias llegaron a tener desvíos promedio relativamente altos. Combinando todos estos

clústeres, solo 22 secuencias alcanzaron desvíos superiores a 0,30 y que llegaron hasta 0,41 (Figura 35). Estos clústeres se diferenciaban en aspectos como su duración, valor neto, cantidad de niveles de volumen y distintos niveles de desvío, pero incluso en los niveles más altos, como el caso con el desvío de 0,41, no se encontraron comportamientos que parezcan ameritar más investigación.

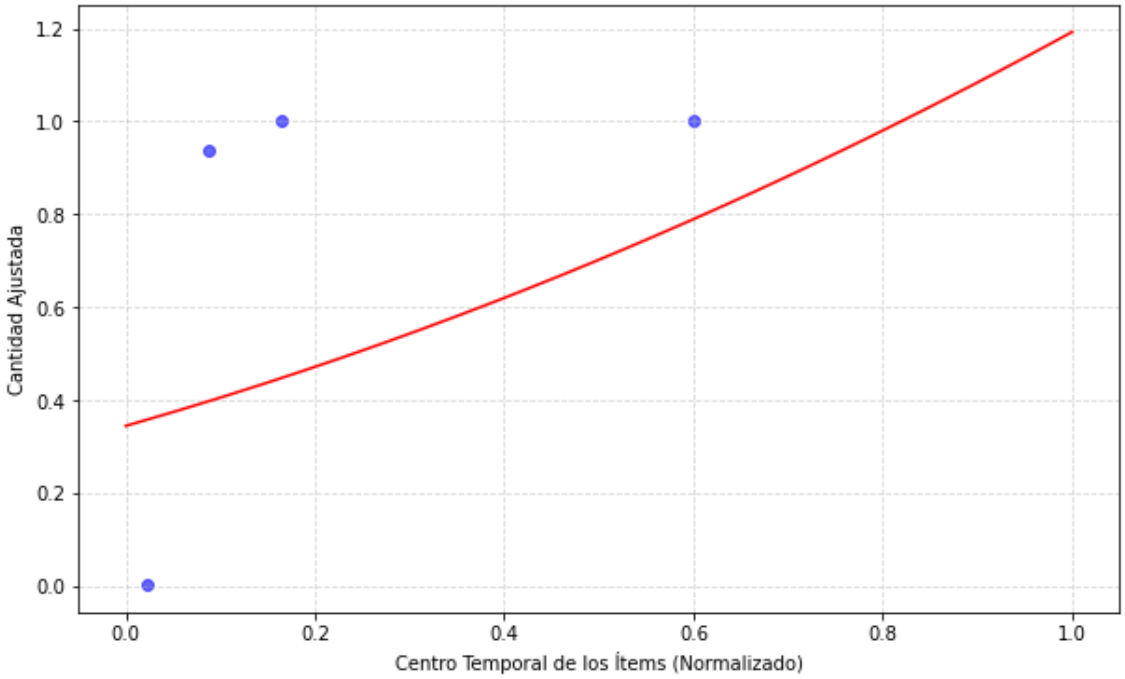

Figura 35. Secuencia con el mayor desvío promedio de los clústeres sin anomalías descubiertas.

Clústeres 3, 9 y 10

Estos clústeres parecen ser una extensión del clúster 0, en el sentido de que todas las secuencias también se mantienen dentro de un mismo nivel de volumen, y que aparecen algunas pocas anomalías aisladas de los mismos tipos que los ya detectados en el clúster 0 (ocho ocurrencias entre todos los clústeres). Estas aparecían entre las secuencias con mayor desvío dentro de cada clúster. En el clúster 3, el desvío promedio máximo fue de 0,34, en el clúster 9, de 0,37 y en el 10, de 0,54. Cuando se revisaban los casos con desvíos menores, ya no se encontraron más anomalías.

Clúster 6

Este grupo se caracteriza por contener algunas secuencias con ítems apilados (con la misma fecha inicial y final), aunque con cantidad constante cuando se contempla el total. Debido a que la solución no se desarrolló como para que pueda procesar ítems apilados, los desvíos que

calcula son altos. El principal motivo es que el procesamiento secuencial de ítems acumula las duraciones. Recordemos esta fórmula para calcular los centros temporales de cada ítem:

$$C_i = S_i + \frac{D_{normalizada_i}}{2}$$

En el ejemplo de la figura 36, los dos primeros puntos en realidad corresponden al mismo período temporal, pero la acumulación hace que parezca que son sucesivos.

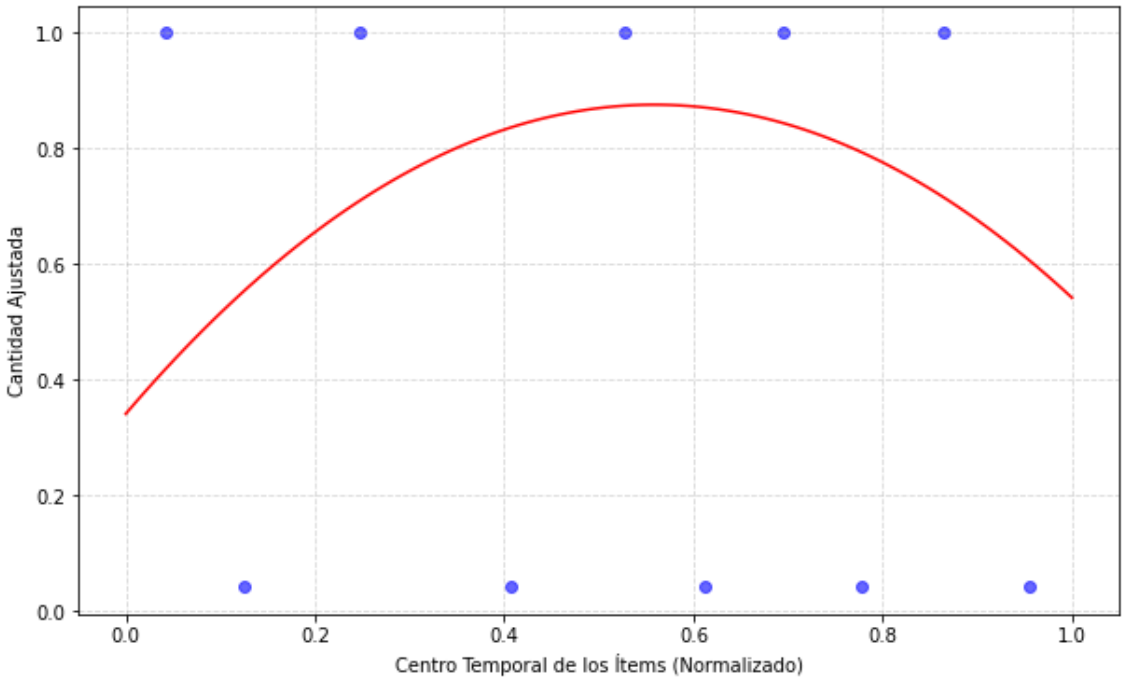

Figura 36. Secuencia con ítems apilados, con cantidad real constante.

Si bien la cantidad real total es constante, sigue llamando la atención por qué los contratos de este clúster se estructuraron de esta forma, lo que es algo que se podría investigar.

Clúster 7

Este clúster agrupa los casos *Hockey Stick* con un alto grado de acierto. Todos los casos que se habían etiquetado manualmente correspondientes a contratos firmados en 2024 quedaron asignados a este clúster, a pesar de que la variable *Hockey Stick* no fue incluida en ninguna de las dos etapas de la solución. Pero además de ellos, se asignaron muchos otros casos que no se habían capturado en el etiquetado manual por no cumplir con el criterio que se había definido de que haya un incremento de nivel de volumen durante el último 10% de la duración del contrato. Sin embargo, estos otros casos cumplen con la característica esencial del grupo, por lo que tiene sentido que haya quedado asignados a este clúster, lo que también es útil para identificar estos casos de forma más fácil y rápida en comparación a como el proceso se había

hecho sin este método. Por ejemplo, en la figura 37 se muestra un caso etiquetado manualmente, en el que podemos ver que la duración de la última fase es menor a la de la figura 38, la cual fue descubierta por el algoritmo de agrupamiento no supervisado, y si bien no posee un desvío demasiado alto (0,36) a comparación de los valores de desvío en todos los clústeres, sí es uno de los desvíos más altos dentro de este clúster.

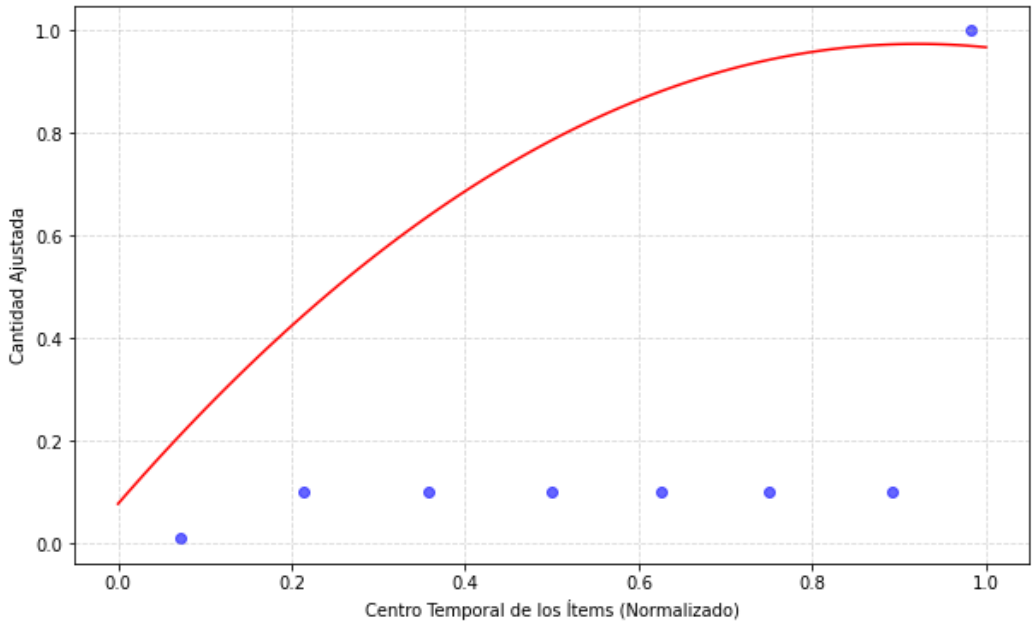

Figura 37. Secuencia *Hockey Stick* que ya había sido etiquetada manualmente.

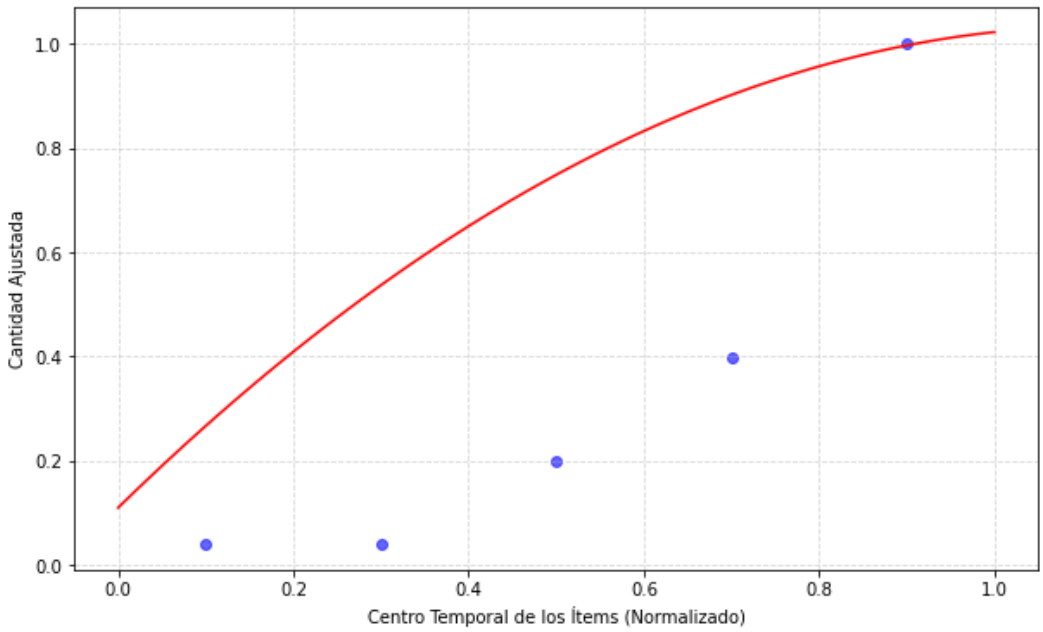

Figura 38. Secuencia *Hockey Stick* descubierta que no había sido etiquetada manualmente.

Aquí solo vimos dos ejemplos, pero en el clúster hay una notable cantidad de casos como estos (se pueden contar en decenas) que en esencia tienen el mismo comportamiento con distinta magnitud de salto de cantidad (y consecuentemente de nivel de volumen) y con distinta duración de la última fase. A medida que uno observa casos con menor desvío promedio, la magnitud de los saltos tiende a ser menor y la duración de la última fase, mayor. Es difícil establecer un umbral fijo para separar lo que es aceptable de lo que no. De hecho, el negocio lo hizo estableciendo la regla de que no puede haber un aumento de nivel de volumen en el último año de contrato, salvo que esto sea aprobado por revisores adicionales. Sin embargo, ¿si el salto de nivel de volumen sucede antes de los últimos trece meses (como en la figura 38), realmente debería ser un comportamiento aceptable? ¿Tiene sentido que estos casos pasen como normales durante el proceso de aprobación? Dado el gran impacto que el nivel de volumen tiene en los ingresos, la regla actual es algo que podría ser revisado. Quizás la magnitud del salto de cantidad es algo que también debería ser considerado en la regla.

En este clúster también se encontró una secuencia "infiltrada" con mayor cantidad al principio del contrato, similar a los casos encontrados en el clúster 0. Sin embargo, como en este caso sí hay un cambio de nivel de volumen, tiene sentido que la secuencia haya sido asignada a este clúster, ya que a fines prácticos, el efecto en términos de reducción de ingresos por alcanzar un nivel de volumen más alto termina siendo el mismo a que si el aumento de nivel de volumen sucede al final del contrato.

Otra variante que se encontró fue una secuencia que además de tener un salto de cantidad al final se la suscripción, también tenía otro salto alrededor del medio (figura 39). En este caso, la duración total de la secuencia es 365 días (la cual ya es inferior a la normal), mientras que cada uno de los saltos tiene una duración de 30 días. Estos saltos permitieron obtener para los 365 días el precio por unidad menor correspondiente al nivel de volumen 4, a pesar de que la mayor parte del tiempo el cliente se encontraba en el nivel 2.

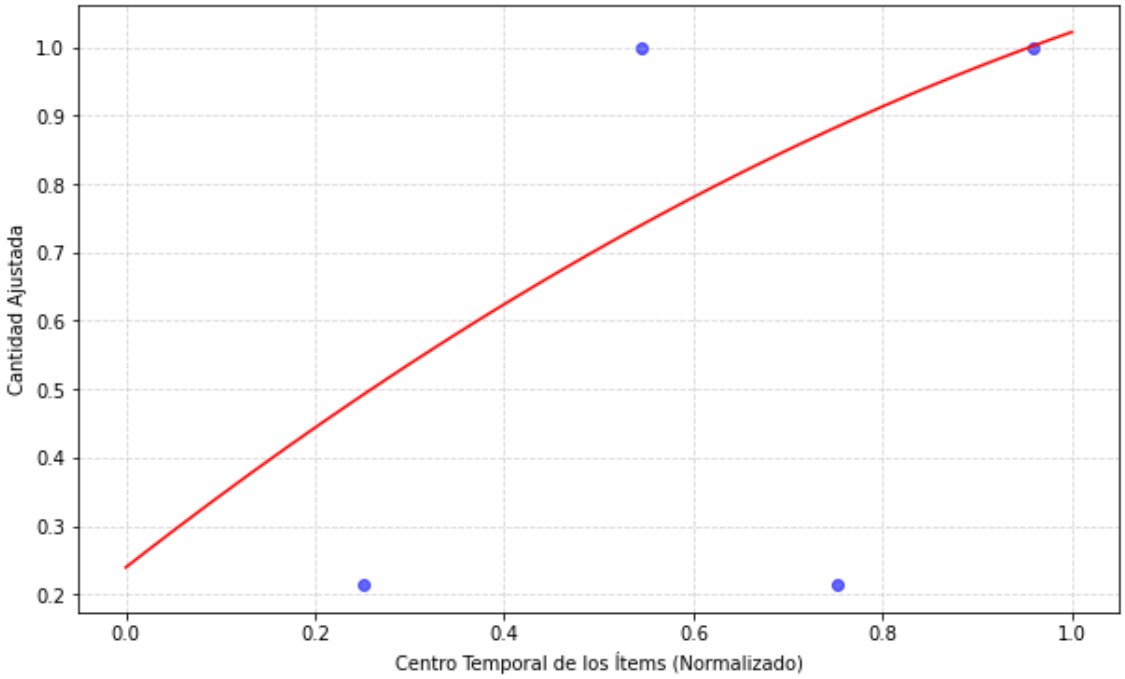

Figura 39. Secuencia con dos breves saltos de cantidad.

Clúster 11

De este grupo lo que llama la atención es la corta duración de las secuencias, la cual es menor a 1.000 días en muchos casos y llegando a un mínimo de 304 días. Teniendo en cuenta que la directiva general del negocio es que los contratos tengan una duración mínima de dos años o 730 días, los motivos detrás de este comportamiento deben ser revisados.

Resumen

En la siguiente tabla (Figura 40) se resumen los resultados de todos los clústeres para obtener una mejor visión general. También se muestran algunos datos complementarios.

Clúster	Cantidad de Secuencias y Porcentaje del Total	Cantidad de Secuencias Anómalas entre las 10 con mayor Desvío Promedio	Cantidad de Niveles de Volumen Distintos	Duración Media (Días)	Media de los Desvíos Promedio	Desvío Promedio Máximo	Tipos de Anomalías Encontradas / Comentarios
0	122 (8%)	8	1	1.708	0,23	0,69	Cantidad decreciente al principio, cantidad decreciente hacia el final, fuerte incremento de cantidad hacia el final, cantidad creciente al principio y decreciente hacia el final, cantidad decreciente a lo largo de toda la duración.
1	159 (10%)	0	2 a 3	1.819	0,15	0,35	N/A
2	193 (12%)	0	2 a 3	2.024	0,19	0,41	N/A
3	86 (5%)	2	1	2.361	0,18	0,34	Cantidad decreciente hacia el final
4	4 (0%)	0	1 a 2	1.402	0,21	0,24	N/A
5	1 (0%)	0	1	2.551	0,08	0,08	Contrato con alto valor neto y larga duración, sin ninguna otra anomalía
6	15 (1%)	10	1	4.763	0,29	0,49	Ítems apilados hacen que parezca que la duración es mucho mayor de lo que realmente es y también hacen que el desvío calculado sea alto, cuando en realidad la cantidad suscripta es constante a lo largo de toda la duración del contrato.
7	187 (12%)	7	2 a 5	1.566	0,16	0,54	Hockey Stick, cantidad decreciente al principio con cambio de nivel de volumen, dos saltos breves de cantidad con cambio de nivel de volumen.
8	176 (11%)	0	1	1.952	0,14	0,28	N/A
9	221 (14%)	1	1	1.953	0,16	0,37	Cantidad decreciente hacia el final
10	105 (7%)	4	1	2.007	0,28	0,54	Cantidad decreciente al principio, cantidad decreciente hacia el final. cantidad creciente al principio y decreciente hacia el final.
11	152 (10%)	0	1	1.042	0,15	0,38	Llama la atención la duración corta.
12	168 (11%)	0	2 a 3	1.495	0,17	0,34	N/A
13	1 (0%)	0	2	1.825	0,09	0,09	Contrato con alto valor neto, sin ninguna otra anomalía

Figura 40. Resumen del análisis de cada clúster.

Por otra parte, en la figura 41 se muestra una representación gráfica de los clústeres luego de aplicar el algoritmo *t-SNE* para reducir la dimensionalidad a dos componentes. Lo más notable es que el clúster 7 que engloba a los casos *Hockey Stick* tiende a encontrarse en la esquina inferior izquierda, mientras que el clúster que engloba a los casos con corta duración tiende a estar en la esquina inferior derecha. También puede notarse que hay un corte que separa los clústeres en dos grandes grupos. En el grupo de la derecha se ubican los otros clústeres que tenían anomalías, aunque tienden a estar bastante centrados. Algo que también puede observarse es que los clústeres que tienen anomalías tienden a ubicarse en la parte inferior del gráfico, a excepción del clúster 10, por lo que podría argumentarse que valores inferiores en el segundo componente de t-SNE podrían indicar mayor probabilidad de encontrar anomalías.

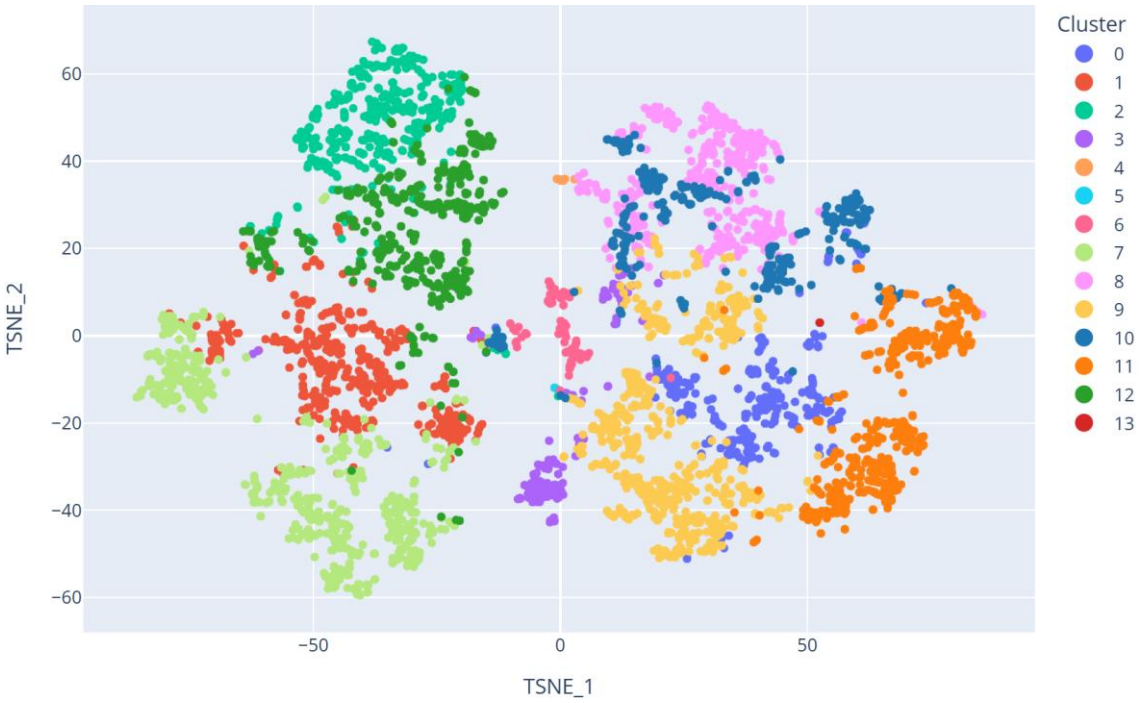

Figura 41. Visualización de los 14 clústeres mediante t-SNE.

3.3.1.2 Segunda Variante: Optimizando K-Medias

3.3.1.2.1 Detalles de Implementación

Un primer aspecto que se podría intentar mejorar de la primera variante es la determinación de *k*, ya que el *Método del Codo* se presta a ambigüedades. Por este motivo, se utilizó el método de Silhouette, que por un lado considera qué tan coherente es cada punto con su clúster mediante la distancia media entre el punto y todos los otros puntos dentro del clúster *(*$a(i)$). Por el otro, mide la separación entre clústeres mediante la distancia media entre un punto y todos los puntos del clúster más cercano $(b(i))$. Para cada punto *i* el Coeficiente de Silhoutte *s(i)* se define como:

$$s(i) = \frac{b(i) - a(i)}{(max(b(i), a(i))}$$

La fórmula normaliza la diferencia entre $b(i)$ y $a(i)$ dividiéndola por el mayor de los dos, asegurando que el resultado esté entre −1 y 1:

$s(i) \approx 1$: El punto está bien agrupado (lejos de otros clústeres, cerca del suyo propio).

$s(i) \approx 0$: El punto está cerca del límite entre clústeres (distancia similar a su propio clúster y al clúster más cercano).

$s(i) < 0$: El punto podría estar en el clúster equivocado (más cerca de otro clúster que del suyo propio).

Una vez que se calcula *s(i)* para todos los puntos y para distintos valores de *k*, se puede calcular un coeficiente promedio, que podría utilizarse para la selección del valor óptimo de *k.* Los resultados se muestran en la figura 42. Allí puede verse que el valor más alto se obtiene con *k* = 2.

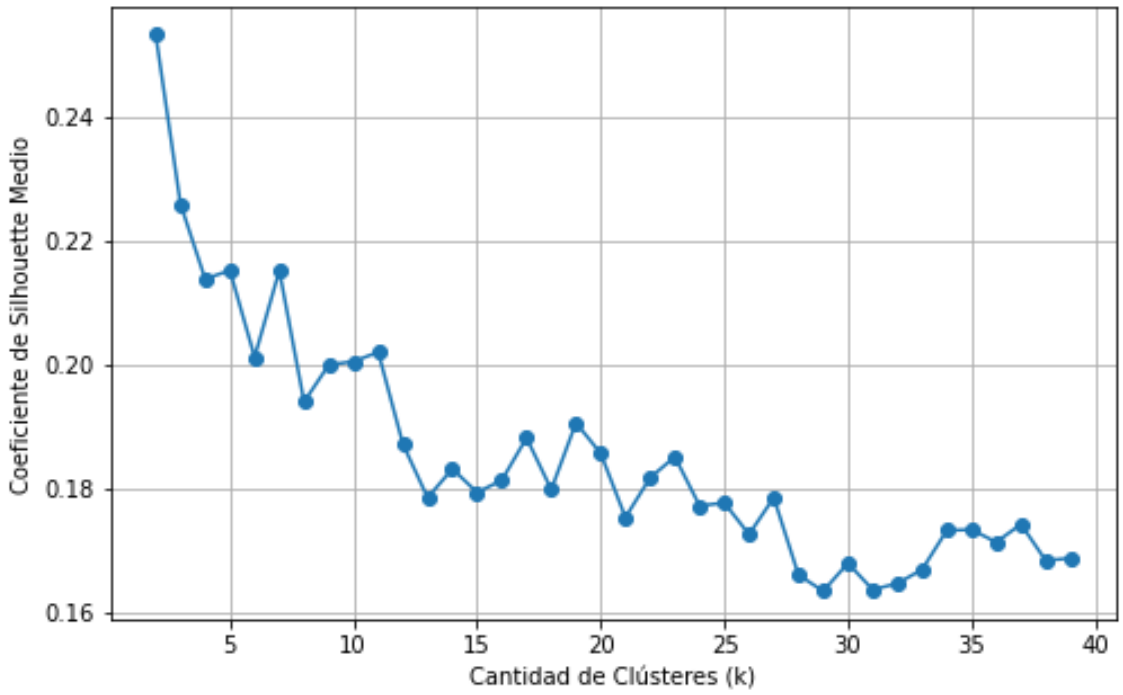

Figura 42. Análisis de Sihouette para la obtención de *k* óptimo (16 variables).

Éste es un resultado que no es útil para el propósito de identificar distintos tipos de anomalías. En general, mediante el método de Silhouette se espera que el valor óptimo de *k* sea mayor a dos debido a que mientras menor sea *k*, mayor sería la distancia intra-clúster. Un motivo para este resultado puede ser la relativa alta dimensionalidad que se está manejando (16 variables), en la que algunas variables pueden estar correlacionadas. Por este motivo, a continuación se aplicó Análisis de Componentes Principales como técnica de reducción de dimensionalidad. Recién con una reducción a tres componentes principales, el método de Silhouette arrojó un óptimo superior a *k* = 2, el cual fue de *k* = 4 con *s* = 0,3553 (figura 43). En la figura 44 se muestran los gráficos del análisis de Silhouette para cuatro y para cinco componentes principales, los cuales arrojan el infructuoso óptimo en *k* = 2. La proporción acumulada de la varianza explicada por los tres componentes es de 0,6269, lo que implica que si bien gran parte de la varianza de

los datos originales es explicada mediante estos tres componentes (62,69%), queda una significativa parte de la información que se está perdiendo (37,31%).

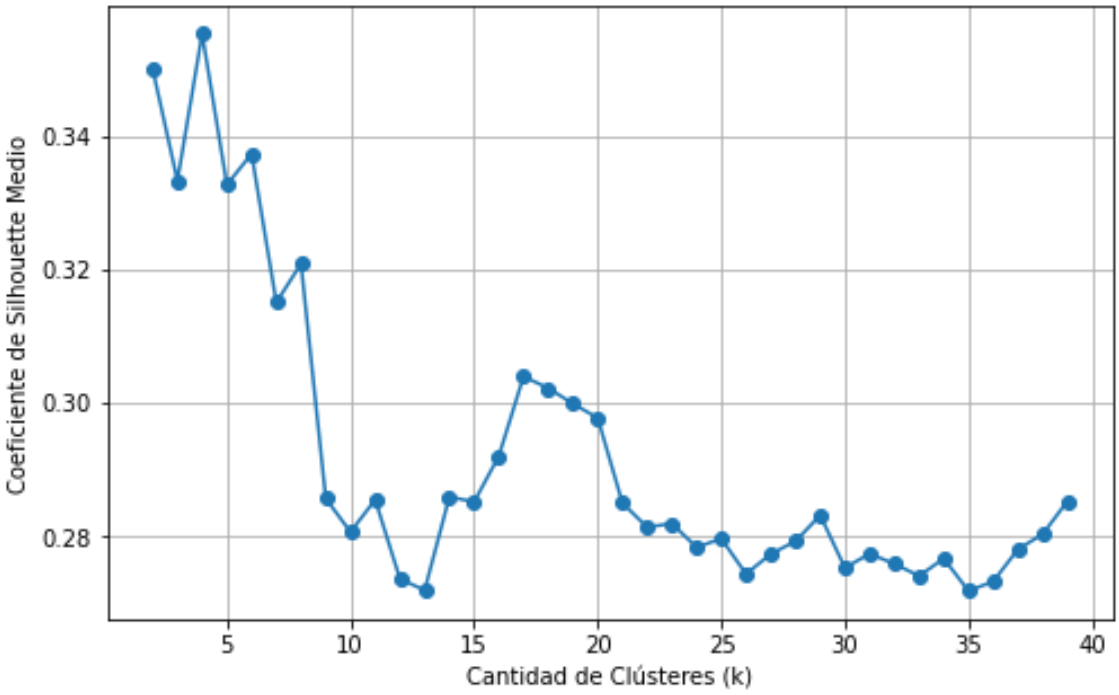

Figura 43. Análisis de Sihouette para la obtención de *k* óptimo con Análisis de Componentes Principales (3 componentes).

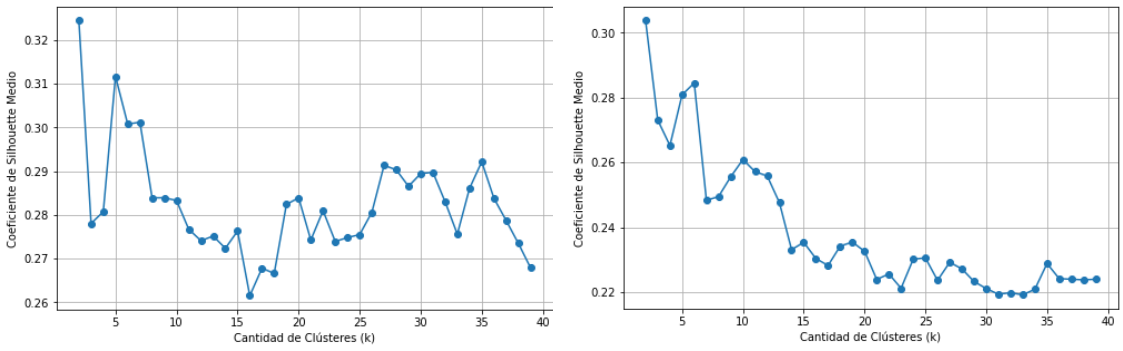

Figura 44. Análisis de Sihouette para la obtención de *k* óptimo con Análisis de Componentes Principales (izquierda cuantro componentes, derecha cinco componentes). Ambos arrojando un óptimo en $k = 2$.

También se realizó el mismo ejercicio, pero utilizando el algoritmo de *k*-medias bisectado, en lugar de su versión tradicional. Este funciona empezando con todos los datos como un único clúster y luego dividiéndolo repetidamente en dos, seleccionando el clúster más grande o menos compacto para cada nueva división, hasta alcanzar el número deseado de clústeres. Como

principales ventajas, puede mejorar la calidad del agrupamiento no supervisado, pudiendo adaptarse mejor a conjuntos de datos complejos, permitiendo divisiones más naturales y obteniendo clústeres con tamaños más balanceados. Además, tiene menor sensibilidad al punto de inicialización, lo que reduce la dependencia de la elección inicial de centroides, disminuyendo el riesgo de resultados subóptimos. Al hacer el análisis de Silhoutte para esta versión, el óptimo (s = 0,3548) también se obtuvo con k = 4 (figura 45). El resultado de s es prácticamente igual, aunque levemente inferior al de *k*-medias tradicional, por lo que se optó por seguir trabajando con la versión tradicional.

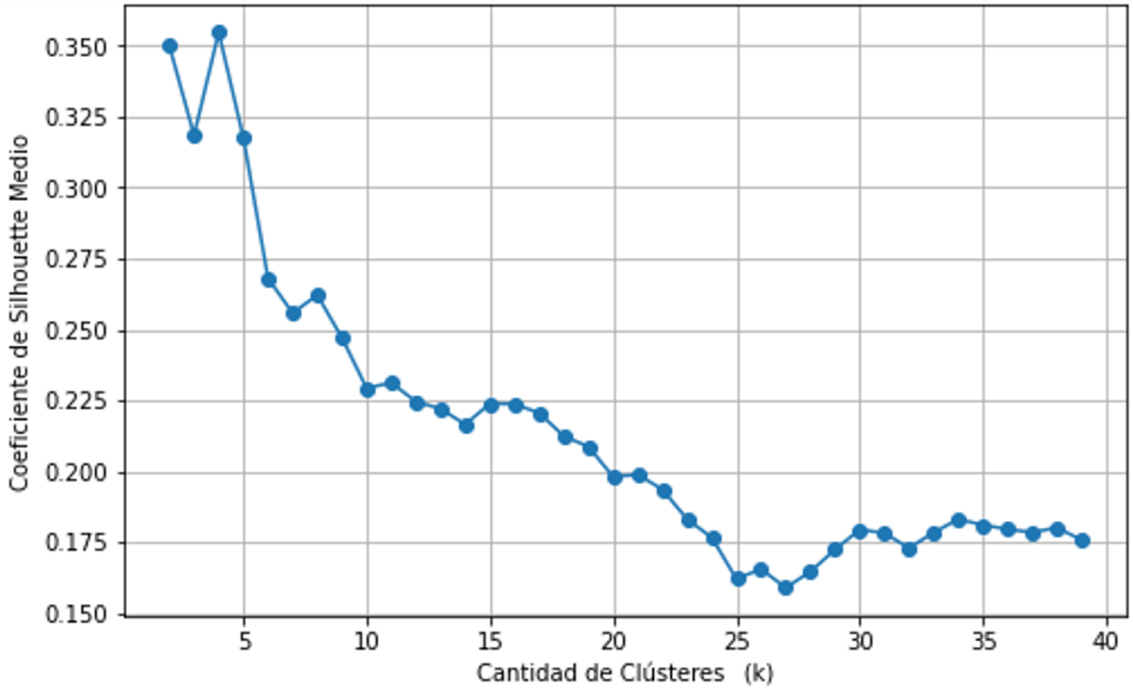

Figura 45. Análisis de Sihouette para la obtención de *k* óptimo con Análisis de Componentes Principales (*k*-medias bisectado).

3.3.1.2.2 Análisis de Resultados

Tal como en la sección 3.3.1.1.2, se combinan los resultados de la etapa de regresión con esta etapa de agrupamiento no supervisado y se obtiene una lista de secuencias ordenada de mayor a menor desvío promedio solo que en vez de catorce clústeres ahora hay cuatro. Ya habiendo identificado secuencias anómalas con la primera variante, éstas se pueden mapear mediante el identificador Cotización-*SKU* a los resultados de esta segunda variante, facilitando el análisis. También se realizó una revisión manual de los resultados para verificar si la nueva forma de agrupar los datos facilita el descubrimiento de nuevos tipos de anomalías. Sin embargo, lo que se obtuvo fue una versión más condensada de la primera variante, en la que no se descubrieron

nuevos tipos de anomalías. Esto suena razonable, ya que pasando de catorce a cuatro clústeres es esperable que no se pueda alcanzar el mismo nivel de granularidad. La ventaja de esta variante radica en que los resultados se pueden mostrar de forma más comprimida, pudiendo eliminar ruido con menor intervención de un analista, el cual en vez de tener que revisar catorce clústeres, sólo tendría que revisar cuatro. En la figura 46 se muestra una tabla con el resumen de los resultados.

Clúster	Cantidad de Secuencias y Porcentaje del Total	Cantidad de Secuencias Anómalas entre las 10 con Mayor Desvío Promedio	Cantidad de Niveles de Volumen Distintos	Duración Media (Días)	Media de los Desvíos Promedio	Desvío Promedio Máximo	Tipos de Anomalías Encontradas / Comentarios
0	381 (24%)	5	1 a 2	1834	0,19	0,54	Cantidad decreciente al principio, cantidad decreciente hacia el final, cantidad creciente al principio y decreciente hacia el final.
1	359 (23%)	0	1 a 4	1853	0,18	0,41	N/A
2	357 (22%)	8	1 a 5	1576	0,15	0,54	Hockey Stick, cantidad decreciente al principio con cambio de nivel de volumen, dos saltos breves de cantidad con cambio de nivel de volumen
3	493 (31%)	10	1 a 2	1895	0,18	0,69	Cantidad decreciente al principio, cantidad decreciente hacia el final, fuerte incremento de Cantidad hacia el final, cantidad creciente al principio y decreciente hacia el final, cantidad decreciente a lo largo de toda la duración, ítems apilados, contrato con alto valor neto, sino ninguna otra anomalía

Figura 46. Resumen de resultados reagrupados.

Se puede ver que ahora el clúster con mayor variedad de tipos de anomalías es el número 3, el cual podría considerarse similar al clúster 0 de la primera variante, aunque ahora también incluye casos como los de los ítems apilados o un contrato con valor notablemente superior al resto. Después, el clúster 0 de esta segunda variante puede verse como una extensión del clúster 3, ya que contiene unas pocas ocurrencias de anomalías de tipos contenidos en el clúster 3.

El clúster 2 parece ser el análogo al clúster 7 de la primera variante. También contiene todos los casos *Hockey Stick*, incluyendo todos los etiquetados manualmente de antemano.

Finalmente, en el clúster 1, el cual tiene desvíos medios y máximos notablemente inferiores a los otros clústeres, no se encontraron anomalías. Esto permitiría afirmar con un alto grado de confianza que cualquier secuencia asignada a este grupo no presenta comportamientos a investigar, mientras que en todos los otros clústeres sí, dependiendo de su grado de desvío.

Una consideración de esta segunda variante de agrupamiento no supervisado respecto a la primera, es que si bien los resultados parecen mostrarse de forma mucho más sintética, a la hora de explorar los datos, la primera variante puede ser de más ayuda por su mejor granularidad. Por ejemplo, los casos de ítems apilados o los de corta duración quizás no se

hubiesen detectado de haber utilizado únicamente la segunda variante, ya que se encontraban más abajo en la lista de secuencias de su clúster, mientras que en la primera variante tenían un clúster dedicado.

Tal como se hizo con la primera variante, a continuación (figura 47) se muestra una visualización de los clústeres en dos dimensiones mediante t-SNE. Allí también puede verse un corte vertical que separa los clústeres en dos. Allí es interesante ver como los clústeres 3 y 0 quedaron agrupados del lado derecho, siendo que se había mencionado que el clúster 0 parecía ser una extensión del clúster 3. Esta visualización podría estar soportando la idea de que estos clústeres tienen características similares.

Del lado izquierdo se encuentra el clúster 1, que no contiene anomalías detectadas, y el clúster 2, que engloba a los casos *Hockey Stick* y sus variantes. Si bien este clúster tiene bastante contacto con el clúster 1, lo que en cierta manera tiene sentido ya que las secuencias *Hockey Stick* tienden a seguir el mismo comportamiento ascendente que las secuencias normales, aunque de forma más pronunciada hacia el final, puede separase notoriamente del clúster 1, ubicándose por encima del mismo.

Tal como en la versión de 14 clústeres, también aquí puede notarse cierta tendencia a que los casos anómalos se encuentren en la parte inferior. En el grupo izquierdo vimos que los casos *Hockey Stick* se encuentran debajo de los que no tienen anomalías, y del lado derecho, el clúster 3 que contiene más anomalías que el 0, también se ubica por debajo. Esto da soporte a la idea de que valores inferiores en el segundo componente de t-SNE se relacionan con mayor probabilidad de encontrar anomalías.

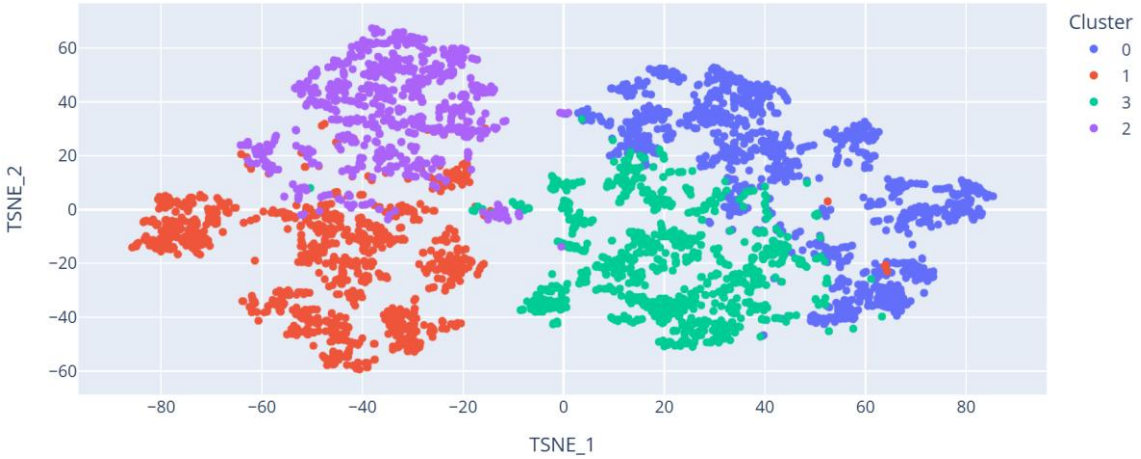

Figura 47. Visualización de los 4 clústeres mediante t-SNE.

3.4 Validación

Luego de realizar el trabajo descripto hasta ahora, se obtuvieron datos más recientes correspondientes al último trimestre de 2024 (el análisis que se había hecho contenía datos hasta el tercer trimestre). Estos datos permitieron efectuar una prueba para validar cómo se comporta la solución con datos que no había visto antes. Para ello, lo que se hizo fue agregar los nuevos datos a los cuales con los que ya se había trabajado y ejecutar todo el código. Se utilizó la segunda variante de agrupamiento no supervisado para poder ver los resultados de forma más concisa.

En los nuevos datos, los resultados fueron muy similares a lo que ya se había visto:

- Uno de los clústeres se mantuvo sin anomalías
- Uno siguió agrupando las secuencias *Hockey Stick* y sus variantes similares
- Un clúster agrupó la mayor parte de las anomalías. Y de hecho, se encontró un nuevo tipo en el que la cantidad suscripta cae alrededor de la mitad y luego vuelve a aumentar por encima de la cantidad inicial (véase figura 48).
- El clúster restante contenía algunas anomalías aisladas, complementando a las de la línea anterior.

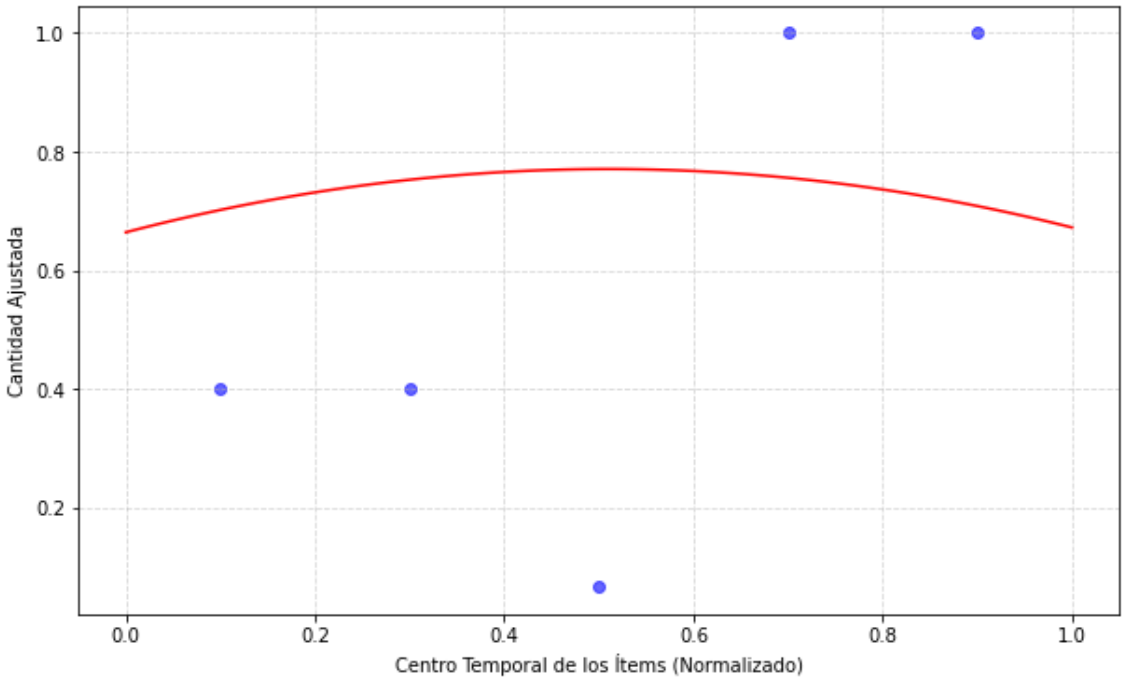

Figura 48. Secuencia con disminución y posterior aumento de cantidad.

En conclusión, se pudo validar que con datos nuevos la solución se comporta de la misma manera que como durante su desarrollo y entrenamiento inicial, teniendo incluso la capacidad

de seguir descubriendo nuevos tipos de anomalías, lo que verifica el poder de generalización de la solución.

4 Resultados

4.1 Desempeño de la Solución

En la sección 3.1 se habían definido los siguientes criterios de evaluación de éxito:

(1) Facilidad para detectar anomalías de tipos ya conocidos (*Hockey Stick*).
(2) Capacidad de detectar tipos de anomalías desconocidos hasta el momento.
(3) Injerencia de falsos positivos y falsos negativos

A continuación se procederá a evaluar la solución propuesta en base a estos criterios.

En cuanto a la facilidad de detectar anomalías de tipos ya conocidos, la solución tuvo un muy buen desempeño, ya que en ambas variantes logró agrupar en un único clúster todos los casos *Hockey-Stick*, que es el único tipo de anomalía conocido hasta el momento de hacer este trabajo. Cabe destacar que además de asignar a este clúster el 100% de los casos de 2024 que se habían etiquetado manualmente, la solución también asignó gran cantidad de casos a este grupo, que en esencia tenían el mismo comportamiento, aunque con distinto grado de profundidad o severidad. Para su evaluación, el cálculo del desvío promedio fue de gran ayuda, dando a un analista una estimación preliminar de qué tan severo puede ser el efecto en cada caso, brindándole la posibilidad de enfocarse en los casos con mayor desvío.

También debe considerarse que para la identificación de casos en el etiquetado manual (y en el análisis que se había hecho luego del descubrimiento real de este comportamiento), sólo se había considerado la duración de la fase con mayor cantidad, con el criterio de que sea menor al 10%. En cambio, la solución propuesta no sólo contempla la temporalidad de la fase con mayor cantidad, sino que también la magnitud del salto de cantidades entre fases, lo que también es capturado en el desvío promedio. Esto también explica cómo puede haber casos que se habían etiquetado manualmente como *Hockey Stick* que resultaron en desvíos promedio muy bajos (véase la figura 27). Éstos tenían saltos de cantidad más sutiles, indicando que la potencial pérdida de ingresos por diferencia de niveles de volumen es pequeña.

Por lo contrario, hay muchos casos detectados por esta nueva solución como el de la figura 38, que si bien no cumplen con el criterio del 10%, ni con el criterio de un año que estableció el negocio, tienen un potencial de disminución de ingresos superior debido a su salto de nivel de

volumen más marcado, por lo que no sería apropiado ignorarlos. Contemplando todo esto, se puede argumentar que la nueva solución tiene un mejor desempeño para detectar casos *Hockey Stick* con mayor potencial para causar pérdidas de ingresos significativas para el negocio.

Respecto a la capacidad de detectar nuevos tipos de anomalías, mediante la solución propuesta se pudo llegar a descubrimientos razonables, descubriendo comportamientos que probablemente no se hubiesen encontrado sin esta solución, ya que en la compañía no hay un proceso sistematizado para ello. En concreto, se identificaron los siguientes tipos de secuencias anómalas, de los cuales muchos de ellos podrían conllevar la pérdida o falta de realización de ingresos potenciales:

- Cantidad decreciente cerca del inicio del contrato
- Cantidad decreciente al final del contrato
- Incremento de cantidad al final del contrato (en *SKUs* con un único nivel de volumen)
- Cantidad creciente al principio y decreciente hacia el final, alcanzando la cantidad máxima cerca del medio.
- Cantidad decreciente a lo largo de toda la duración del contrato
- Dos saltos breves de cantidad con cambio de nivel de volumen
- Ítems apilados
- Corta duración total
- Disminución y posterior aumento de cantidad

Teniendo en cuenta que el negocio actualmente no cuenta con un proceso sistematizado para detectar este tipo de casos, los resultados obtenidos aquí pueden ser una contribución significativa para continuar con futuras investigaciones en cuanto a las implicancias de estos comportamientos. Tener un método como este, que brinde transparencia y visibilidad sobre nuevos tipos de casos anómalos es el primer paso para evaluar si hace falta implementar nuevas reglas de negocio, si estos comportamientos no constituyen ningún riesgo y consecuentemente no son necesarias más acciones, o si existe la oportunidad de diseñar iniciativas para optimizar la generación de ingresos. Una primera aproximación para intentar responder estas cuestiones se hace en la sección 4.3 (Análisis Prescriptivo).

También debe considerarse que con el desarrollo de esta solución, ya se tiene un script de código que puede volver a correrse en el futuro con datos más actualizados con el fin de monitorear la aparición de nuevos tipos de anomalías no detectados en este trabajo.

En lo concerniente a la injerencia de falsos positivos y falsos negativos llegamos a que este es uno de los principales aspectos que se podrían y deberían mejorar. Una vez que el código de la

solución corre, el resultado es un archivo con todos los pares Cotización-*SKU* asignados a clústeres y ordenados de mayor a menor desvío promedio. Este ordenamiento, como ya mencionamos, es de gran ayuda para identificar los casos que realmente son anómalos, pero entre estos casos hay muchos que en realidad no tienen nada que llame la atención cuando se los revisa en detalle individualmente, incluso en los clústeres con mayor cantidad de anomalías. De la misma manera, hay casos que sí tienen aspectos a investigar en más detalle, que tienen desvíos bajos. Es decir, con el método actual es muy difícil establecer un corte para poder decidir de forma precisa si un caso es anómalo o no. ¿Cómo se podría mejorar esto? Una primera idea es darle mayor importancia al valor neto a la hora de seleccionar los casos a investigar. Se podría generar una métrica combinada entre el desvío y el valor neto, de forma que la relevancia de los desvíos sea influenciada por su valor neto. Es decir, se podría ponderar el desvío por el valor neto de la secuencia. Si bien esta idea permitiría enfocarse con mayor facilidad en casos con más relevancia inmediata en términos monetarios y con menor tiempo invertido por el analista, se corre un riesgo mayor de pasar por alto comportamientos que todavía no aparecen en contratos grandes, y que cuya ocurrencia podría estar en aumento.

Resumiendo esta evaluación de desempeño, lo positivo es que la solución fue exitosa en la identificación de anomalías de tipo ya existente y en la detección de nuevos tipos de anomalías, lo que ni siquiera era posible antes. Además, tanto para los casos *Hockey Stick* como para los nuevos descubrimientos, ahora se cuenta con el código que se puede volver a ejecutar en el futuro con datos actualizados y mínimos ajustes para seguir monitoreando la aparición de anomalías, lo que constituye una gran reducción de esfuerzo a comparación de tener que hacer un trabajo en planillas de cálculo a mano desde cero con los nuevos datos. Lo que es un aspecto a mejorar es la eficiencia del discernimiento entre casos anómalos y normales, pero ya se cuenta con ideas sobre cómo se podría mejorar este aspecto.

4.2 Implementación en el Entorno Real

Dado que la solución propuesta ya tiene la capacidad de aportar valor mediante la identificación de anomalías de tipo conocido y de descubrir nuevos tipos de anomalías, podría implementarse en el entorno real de la compañía tal como se hizo en la sección 3.4 (Validación). Es decir, luego de un período regular de tiempo que puede ser mensual, trimestral, o semestral, se pueden agregar los datos nuevos al conjunto de datos histórico, y el código se puede volver a ejecutar con la totalidad de los datos. Como para los datos históricos ya se había revisado la lista de resultados y se habían marcado los casos anómalos, los mismos resultados se pueden mapear a la lista generada conteniendo también las secuencias nuevas. Como tanto las secuencias nuevas

como las históricas se repartirían en los mismos clústeres, se pueden utilizar los casos marcados de las secuencias históricas para decidir a qué clústeres prestarles especial atención al revisar las secuencias nuevas. Prestándole especial atención a los clústeres que ya tenían anomalías en los datos históricos, y dentro de cada clúster a las secuencias con mayor desvío, las anomalías del nuevo período pueden identificarse con relativa facilidad. Sin embargo, la determinación final sobre si cada secuencia es anómala o no sigue quedando a criterio del analista que revisa la lista entregada al ejecutar el código.

De este modo, si bien el modelo se vuelve a entrenar periódicamente en su totalidad, la revisión de resultados solo se hace para los datos nuevos. Esto permitiría implementar un ciclo como el de la figura 48, mediante el cual el conjunto de datos se va enriqueciendo más con cada nuevo período, lo que asimismo facilita la revisión por parte de un analista.

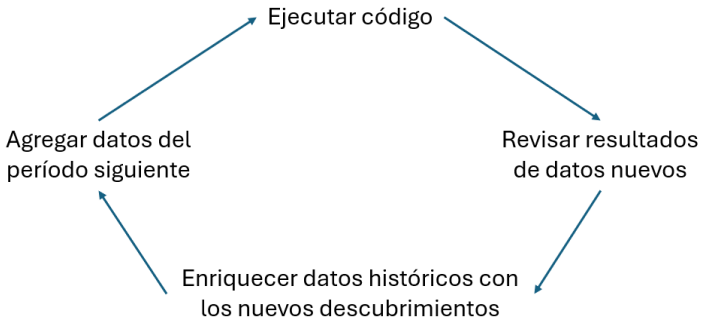

Figura 48. Ciclo de utilización de la solución propuesta.

Una medida de seguridad sería revisar también los clústeres que no presentan anomalías mapeadas de los resultados históricos. Esto puede hacerse fácilmente revisando unas pocas secuencias (por ejemplo, tres) de las que más desvío tienen del clúster en cuestión. Si estas no presentan anomalías, muy probablemente todas los demás del mismo clúster tampoco.

El proceso descripto arriba puede implementarse con cualquiera de las dos variantes propuestas. Si se desea un análisis más riguroso, se recomienda utilizar la primera variante con catorce clústeres. Si se desea un análisis más rápido que aún así tiende a captar la mayoría de los tipos de anomalías relevantes, se puede utilizar la variante con reducción de dimensionalidad y cuatro clústeres.

4.3 Análisis Prescriptivo

En esta sección se procederá a sugerir próximos pasos para aplicar a cada tipo de anomalía que surgió del análisis de resultados.

En el caso de los contratos *Hockey Stick* vimos que durante el período de tiempo para el cual se analizaron los resultados (2024) siguió habiendo una cantidad significativa de casos, a pesar de que ya se había implementado una regla de negocio para evitarlos (aunque se pueden obtener aprobaciones adicionales para proceder con estos casos). Recordemos también que esta regla implementada en el sistema *CPQ* verificaba la presencia de aumentos del nivel de volumen en el último año de duración de las cotizaciones. La presencia de la notable cantidad de casos en 2024 lleva a preguntas como las siguientes: ¿Bajo qué condiciones excepcionales fueron aprobados estos casos? ¿Los aprobadores eran conscientes del impacto en términos de facturación? ¿La regla existente está bien diseñada al considerar sólo el último año de duración? ¿Podría diseñarse una regla mejor? Tal como se propuso en el análisis del clúster 7 en la sección 3.3.1.1.2, a lo mejor se podría implementar una regla en el sistema *CPQ* que verifique la magnitud del salto de volumen entre fases incluso antes del último año del contrato, por ejemplo, podría considerarse la segunda mitad de la duración total. Si en este período la regla detecta un salto de nivel de volumen con un aumento en la cantidad suscripta mayor a cierto umbral (por ejemplo 40%), se podría hacer que el sistema *CPQ* también envíe el caso para la revisión de los aprobadores adicionales. Sin embargo, la determinación de los umbrales y los detalles de implementación de esta regla llevaría un análisis adicional más detallado.

Para el caso de los nuevos tipos de anomalías descubiertos también podrían plantearse preguntas como ¿qué soluciones o *SKUs* suelen tener esos comportamientos? ¿por qué se generan? ¿los clientes realmente necesitan suscripciones con estos comportamientos? ¿hay algún riesgo? Especialmente los que tienen cantidades decrecientes (en todas sus variantes) llaman la atención porque a pesar de que hay reglas para evitar este comportamiento, vimos que está presente. Entonces podemos preguntarnos: ¿la aplicación de las reglas debe hacerse de forma más estricta? ¿o si efectivamente no hay riesgo alguno, quizás se podría ofrecer más flexibilidad a los clientes en general, con suscripciones que permitan cantidades descendientes? ¿Cuál sería el impacto en términos de ingresos, costos, adquisición de clientes, su satisfacción y su retención? Todas estas son preguntas que el negocio debería revisar detalladamente teniendo en cuenta las implicancias estratégicas que conllevan, ya que impactan la forma en que los clientes consumen las soluciones ofrecidas. Sin embargo, se podrían sugerir los siguientes primeros próximos pasos por cada tipo de anomalía descubierto:

Cantidad decreciente cerca del inicio del contrato: La principal sugerencia aquí es tratar de entender el motivo por el que sucede esto. Por ejemplo, identificar en qué tipo de soluciones sucede y bajo qué contexto. Mediante una primera revisión de estos casos se vio que muchos de estos casos se generaban con soluciones de almacenamiento. Esta cantidad extra al comienzo de un contrato quizás es solamente necesaria para los clientes durante ese período para completar un proceso de migración de sistemas. Sin embargo, esto es algo que debe ser confirmado. También podría analizarse si los clientes le pueden dar otro uso a estas soluciones de modo que la cantidad contratada inicialmente pueda extenderse, aumentando la facturación para la empresa y la utilidad para los clientes.

Cantidad decreciente al final del contrato: Lo primero que se sugeriría analizar más detalladamente es nuevamente el motivo por el que se da este comportamiento: ¿Los clientes realmente no esperan utilizar la solución en cuestión hacia el final de su suscripción? De ser así, ¿por qué? Debido a que este comportamiento tendía a impactar únicamente un período corto al final de la suscripción inicial, quizás no tiene gran impacto en los ingresos durante ese período, pero sí podría revisarse más detalladamente el impacto en las renovaciones, dado que en general las mismas se efectúan con las cantidades vigentes al final de la suscripción inicial, lo que sí podría impactar significativamente la facturación durante los períodos de renovación. Es por esto que si se verificara que el motivo de este comportamiento es la falta de utilización de la solución, podrían implementarse programas para incentivarla y que consecuentemente los clientes no reduzcan las cantidades contratadas.

Cantidad decreciente a lo largo de toda la duración del contrato: Se sugiere investigar las mismas preguntas que en el caso de arriba. Si bien el impacto negativo en la facturación ya es más notorio durante el período de suscripción inicial, la casusa subyacente puede ser similar.

Incremento de cantidad al final del contrato (en *SKUs* con un único nivel de volumen): La principal pregunta aquí es: ¿Por qué ese aumento se da tan tarde? ¿Podría adelantarse el uso por parte de los clientes de esos volúmenes de suscripción más altos y aumentar así la facturación? Quizás esto también podría lograrse mediante programas para incentivar y acelerar la adopción.

Cantidad creciente al principio y decreciente hacia el final, alcanzando la cantidad máxima cerca del medio: Mismo aquí, debería investigarse por qué la cantidad máxima se alcanza durante un período relativamente corto cerca del medio de la suscripción y si existe la posibilidad de extender ese período, adelantando su inicio, e idealmente haciendo que la cantidad máxima alcanzada se mantenga hasta el final de la suscripción.

Disminución y posterior aumento de cantidad: Se trata del caso inverso al anterior, pero el tratamiento sería análogo. Debería verificarse el motivo por el cual la cantidad cae y evaluar si en estos casos se puede incentivar la contratación y la adopción de mayor volumen durante el "valle" de cantidad.

Dos saltos breves de cantidad con cambio de nivel de volumen: Se considera que esta es una variante del comportamiento *Hockey Stick*, por lo que su tratamiento debería ser el mismo. Es decir, estos casos deben poder detectarse mediante reglas automatizadas y ante su detección, se debe verificar si los clientes realmente necesitan una suscripción con tal característica.

Ítems apilados: Esta anomalía a priori no parece conllevar ningún riesgo debido a que el total de las cantidades suscriptas se mantenía constante en los casos vistos. Sin embargo, se recomienda investigar el motivo por el cual las cantidades se estructuran de esta manera, por si hay algún aspecto que sí pueda tener algún impacto en el negocio.

Corta duración total: Teniendo en cuenta que la directiva general del negocio es que los contratos tengan una duración mínima de dos años, tendría sentido investigar los motivos por los cuales llegan a ofrecerse suscripciones con menor duración. Si bien ya existen reglas implementadas que alertan a los aprobadores sobre la presencia de esta anomalía, se recomienda verificar bajo qué condiciones estas cotizaciones fueron aprobadas. Esto es de especial importancia debido a que a la hora de determinar los precios de las soluciones ofrecidas hay costos que se distribuyen a lo largo de la duración mínima esperada, y en el caso de que la duración sea menor, existe el riesgo de que la totalidad de los costos no esté cubierta. Esto es algo que también podría analizarse en mayor detalle.

Para todos los tipos, algo que debe tenerse en cuenta es su potencial impacto positivo para que los clientes acepten las cotizaciones. O sea, uno debería preguntarse: De no haber ofrecido la cotización así, ¿el cliente hubiese aceptado la propuesta? ¿Existía riesgo de que elija a un competidor? Quizás al ofrecer las cotizaciones con las anomalías descubiertas se pudo ganar negocios que no se hubiesen podido cerrar de otra manera. Y si bien se vio que podría estar la posibilidad de obtener aún más ingresos, en caso de endurecer las reglas también existe el riesgo de perder ventas si los clientes no están dispuestos a pagar los precios superiores o no necesitan las soluciones durante el tiempo adicional. El balance entre permitir cierta flexibilidad en la estructuración comercial de las suscripciones (permitir ciertas anomalías) y endurecer reglas es un aspecto que debe ser investigado con mayor profundidad y debatido en la compañía, a fin de optimizar los ingresos y el margen de ganancias a nivel agregado.

5 Conclusiones

5.1 Valor Agregado

En la sección 1.2 se había definido el objetivo de "desarrollar un sistema de detección de anomalías que facilite su identificación temprana y permita reducir el tiempo de reacción del área de Control de Gestión para tomar las medidas apropiadas". Luego del desarrollo del método, su implementación y el análisis de los resultados, podemos concluir en que este objetivo ha sido cumplido mediante una solución que aplica una combinación de técnicas de regresión y agrupamiento no supervisado ampliamente conocidas, al dominio específico de este problema. La solución desarrollada permite detectar comportamientos extraños en toda la población de cotizaciones históricas que llevaron a la firma de un contrato, para lo que antes ni siquiera había un proceso establecido. Vale aclarar que dado el gran volumen de datos y que la revisión de los resultados de la solución no fue exhaustiva, puede haber *inliers* que al tener desvíos reducidos, no hayan llegado a la atención del analista. Es decir, hay que tener en cuenta que si bien la solución facilita la identificación de anomalías, no garantiza la detección exhaustiva, tratándose de un modelo no supervisado. Asimismo, la detección exhaustiva tampoco era el objetivo planteado, sino que la toma de conocimiento de distintos patrones anómalos.

La solución puede ser ejecutada regularmente (por ejemplo, cada trimestre), entregando resultados más exhaustivos en el ámbito de las anomalías de tipo conocido *Hockey Stick* y brindando además la posibilidad de explorar anomalías de tipo desconocido, para lo cual no había un proceso en la compañía, lo que facilitaba que no llamen la atención en áreas como Control de Gestión. Si bien hay múltiples aspectos que pueden ser mejorados, especialmente para que la identificación de casos anómalas sea aún más eficiente, esta solución ya sirve como una herramienta para descubrir situaciones que pasarían desapercibidas de no haberse detectado con este medio.

En concreto, esta herramienta sirve para ampliar la perspectiva y la cobertura del área de Control de Gestión, para que pueda tomar conciencia de la existencia de estas situaciones y con ello poder analizarlas en mayor detalle y de ser necesario, debatir con otra áreas del negocio como Ventas, Costos, o la misma Dirección Estratégica, si los comportamientos detectados constituyen algún riesgo, si pueden ser ignorados, o incluso si tienen algún beneficio para el negocio y deberían ser incentivados, optimizando la facturación y mejorando la satisfacción de los clientes. Es decir, sirve para brindar transparencia, plantearse nuevas preguntas y prestarle atención a aspectos a los que quizás no se les hubiese dado importancia.

5.2 Posibles Puntos de Mejora y Próximos Pasos

En cuanto a mejoras de la solución propuesta en este trabajo, ya se mencionó la idea de facilitar la selección de casos anómalos mediante la creación de una métrica combinada entre valor neto y desvío. Por otra parte, sería interesante generar un modelo de regresión similar al que ya se implementó, pero basado en el nivel de volumen de cada fase en vez de en la cantidad. Este modelo no sería tan granular como el que se ha implementado en el sentido de que perdería la información de variaciones de cantidad dentro de un mismo nivel de volumen, pero al mismo tiempo también podría eliminar mucho ruido que podría no aportar valor al análisis. Por este motivo es una opción que se podría explorar y eventualmente combinar con el modelo ya desarrollado.

Por otro lado, sería interesante explorar la posibilidad de utilizar otras técnicas más complejas para la detección de las anomalías. Uno de los desafíos del problema que se abordó es que el largo de las secuencias es variable. Esto limitaba la elección de posibles técnicas a utilizar. Sin embargo, algunas como Redes Neuronales Recurrentes o Transformers con Mecanismos e Atención podrían llegar a aportar puntos de vista diferentes al análisis.

También, dado que con esta solución ya se pudo etiquetar gran cantidad de secuencias de distintos tipos y el etiquetado podría expandirse aún más si la solución se implementa ejecutándola periódicamente, también se podría explorar la posibilidad de desarrollar modelos de aprendizaje supervisado, que tendrían el potencial de descubrir con más precisión y menor dedicación por parte de un analista los tipos de anomalías ya conocidos.

Otra mejora para poder hacer un análisis más exhaustivo sería desarrollar un método para poder convertir los contratos estructurados con fases “Delta” a fases “Full”. Esto permitiría incluirlos en el análisis, haciendo que los resultados sean aún más representativos de toda la población de datos. Finalmente, se podría evaluar la posibilidad de realizar un análisis similar a este, pero para ventas adicionales y renovaciones, en vez de únicamente para ventas iniciales. Sin embargo, esto probablemente requeriría cambios mayores en el modelo, ya que tendría sentido evaluar los contratos vinculándolos a su venta inicial y no aisladamente, como se ha hecho en este trabajo.

Sin importar cuáles sean las mejoras técnicas que se implementen en el modelado, el próximo gran paso para que esta solución pueda materializar el agregado de valor a la compañía, es la investigación interna sugerida en la sección 4.3 (Análisis Prescriptivo). Para los tipos de anomalías nuevos descubiertos se debe debatir si se siguen permitiendo y en el caso afirmativo

bajo qué condiciones, balanceando la satisfacción de las necesidades y expectativas de clientes con la optimización de ingresos de la compañía.

En conclusión, en este trabajo se pudieron cumplir los objetivos que se habían propuesto con una solución que aún tiene gran potencial para seguir mejorándose, desencadenando análisis más profundos y detallados, y agregar aún más valor a la organización.

6 Glosario

CPQ*: *Configure, Price, Quote – en español Configurar, poner Precio, Cotizar.

Estructura *Hockey Stick*: Se refiere a contratos o cotizaciones que tengan una larga duración con poco volumen y un incremento notorio hacia el final de la suscripción con una corta duración, pareciéndose a la forma de un palo de hockey. A lo largo del trabajo también se utilizan las siglas *HS* para identificarlos.

SKU*: *Stock Keeping Unit: Código para identificar un producto o servicio.

***Big Data*:** Conjunto de datos que, por su gran volumen, requieren técnicas especiales de procesamiento.

7 Referencias

Thudumu, S., Branch, P., Jin, J., & Singh, J. (2020). A comprehensive survey of anomaly detection techniques for high dimensional big data. Journal of Big Data, 7(1), Article 42. https://doi.org/10.1186/s40537-020-00320-x

Dhapte, A. (2025). Anomaly detection market size, growth report - 2030. Market Research Future. https://www.marketresearchfuture.com/reports/anomaly-detection-market-5756

DealHub. (n.d.). *Deal structure*. DealHub. https://dealhub.io/glossary/deal-structure/

Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. (2021). Machine learning for anomaly detection: A systematic review. IEEE Access, 9, 78658-78700. https://doi.org/10.1109/ACCESS.2021.3083060

Majd, M., Najafi, P., Alhosseini, S. A., Cheng, F., & Meinel, C. (n.d.). A comprehensive review of anomaly detection in web logs. Hasso Plattner Institute (HPI), University of Potsdam.

McInnes, G. (n.d.). Oracle's own case study experience with CPQ. Oracle Blogs. https://blogs.oracle.com/cx/post/oracles-own-case-study-experience-with-cpq

Bramer, M. (n.d.). Principles of data mining: Second edition. Springer.

Miao, Z. (2024). Financial fraud detection and prevention: Automated approach based on deep learning. Law School, Southeast University, China.

Blessing, M. (2024). Enhancing fraud detection with deep learning: An in-depth analysis. Obafemi Awolowo University.

Jordan, M., Auth, G., Jokisch, O., & Kühl, J. U. (2020). Knowledge-based systems for the configure price quote (CPQ) process: A case study in the IT solution business. Online Journal of Applied Knowledge Management.

Adelakun, B. O., Antwi, B. O., Fatogun, D. T., & Olaiya, O. P. (2024). Enhancing audit accuracy: The role of AI in detecting financial anomalies and fraud. Finance & Accounting Research Journal.

Ramasamy, Venkatraman. (2024). Overview of Anomaly Detection Techniques across Different Domains: A Systematic Review. International Journal of Computational and Experimental Science and Engineering. 10. 10.22399/ijcesen.522.

Shabir, Jawaid. (2024). A Systematic Review of Anomaly detection using Machine and Deep Learning Techniques. Queueing Systems. 83-94.

Bablu, Tarek & Mirzaei, Hossein. (2025). Machine Learning for Anomaly Detection: A Review of Techniques and Applications in Various Domains.

Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLOS ONE. https://doi.org/10.1371/journal.pone.0152173

Ersoy, P. (2023). Anomaly detection: Identifying critical issues to reduce revenue losses. Dataroid. https://www.dataroid.com/post/anomaly-detection-identifying-critical-issues-to-reduce-revenue-losses

Bollu S. (2024). Anomaly Detection of User Behavioral Events in E-commerce Electronics Stores using SVMs. Blekinge Institute of Technology. https://www.diva-portal.org/smash/get/diva2:1888826/FULLTEXT01.pdf

8 Apéndice

8.1 Código

El código Python de la solución propuesta y del análisis descriptivo se encuentra en el siguiente repositorio:

Repositorio de Código

8.2 Detalle de Secuencias Anómalas Representativas

Figura	Descripción	Fecha de Inicio	Fecha de Finalización	Duración (Días)	Cantidad
30	Secuencia con cantidad decreciente al principio del contrato.	01/07/2024	31/12/2024	184	15
		01/01/2025	31/12/2026	730	1
		01/01/2027	30/06/2027	181	1
31	Secuencia con cantidad decreciente hacia el final del contrato.	01/06/2024	31/08/2026	821	350
		01/09/2026	31/05/2027	272	350
		01/06/2027	31/05/2029	730	5
32	Secuencia con fuerte incremento de cantidad hacia el final del contrato.	01/09/2024	31/12/2024	121	3
		01/01/2025	31/08/2025	242	3
		01/09/2025	31/12/2025	121	3
		01/01/2026	31/08/2026	242	3
		01/09/2026	31/12/2026	121	3
		01/01/2027	31/08/2027	242	3
		01/09/2027	31/12/2027	121	3
		01/01/2028	31/08/2028	243	3
		01/09/2028	31/12/2028	121	3
		01/01/2029	31/08/2029	242	51
33	Secuencia con cantidad creciente al principio y decreciente hacia el final.	01/10/2024	31/12/2024	91	14212
		01/01/2025	31/12/2025	364	16546
		01/01/2026	31/12/2026	364	37370
		01/01/2027	31/12/2027	364	29956
		01/01/2028	31/12/2028	365	22528
34	Secuencia con cantidad decreciente a lo largo de toda su duración.	01/09/2024	31/12/2024	121	4001
		01/01/2025	31/08/2025	242	4001
		01/09/2025	31/12/2025	121	4001
		01/01/2026	31/08/2026	242	2000
		01/09/2026	31/12/2026	121	2000
		01/01/2027	31/08/2027	242	1000
		01/09/2027	31/12/2027	121	1000
		01/01/2028	31/08/2028	243	500
		01/09/2028	31/12/2028	121	500
		01/01/2029	31/08/2029	242	100

36	Secuencia con ítems apilados, con cantidad real constante.	01/02/2025	31/01/2026	364	11684
		01/02/2026	31/12/2027	698	11684
		01/01/2028	31/12/2028	365	11684
		01/01/2029	31/12/2029	364	11684
		01/01/2030	31/01/2031	395	11684
		01/02/2025	31/01/2026	364	500
		01/02/2026	31/12/2027	698	500
		01/01/2028	31/12/2028	365	500
		01/01/2029	31/12/2029	364	500
		01/01/2030	31/01/2031	395	500
37	Secuencia *Hockey Stick* que ya había sido etiquetada manualmente.	21/06/2024	20/06/2025	365	10
		21/06/2025	20/06/2026	365	100
		21/06/2026	20/06/2027	365	100
		21/06/2027	20/06/2028	365	100
		21/06/2028	20/03/2029	273	100
		21/03/2029	20/03/2030	365	100
		21/03/2030	20/03/2031	365	100
		21/03/2031	20/06/2031	92	1001
38	Secuencia *Hockey Stick* descubierta que no había sido etiquetada manualmente.	01/10/2024	30/09/2025	365	10
		01/10/2025	30/09/2026	365	10
		01/10/2026	30/09/2027	365	50
		01/10/2027	30/09/2028	365	100
		01/10/2028	30/09/2029	365	251
39	Secuencia con dos breves saltos de cantidad.	01/05/2024	31/10/2024	184	15
		01/11/2024	30/11/2024	30	70
		01/12/2024	31/03/2025	121	15
		01/04/2025	30/04/2025	30	70
48	Secuencia con disminución y posterior aumento de cantidad.	31/03/2025	30/03/2026	365	6
		31/03/2026	30/03/2027	365	6
		31/03/2027	30/03/2028	365	1
		31/03/2028	30/03/2029	365	15
		31/03/2029	30/03/2030	365	15