UNIVERSIDAD DE ANTIOQUIA

DESARROLLO DE SOLUCIÓN ANALÍTICA PARA LA PREDICCIÓN DE LA DEMANDA DE LÍNEA

Autor:

Juan Esteban Cano Ospina

Universidad de Antioquia

Facultad de Ingeniería, Departamento de Ingeniería Industrial

Medellín, Colombia

2021

Desarrollo de solución analítica para la predicción de la demanda de línea

Juan Esteban Cano Ospina

Informe de práctica como requisito para optar al título de:

Ingeniero Industrial

Asesor:

Carlos Mario Llano Ortiz

Ingeniero mecánico

Universidad de Antioquia

Facultad de Ingeniería, Departamento de Ingeniería Industrial.

Medellín, Colombia

2021

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

RESUMEN

Una de las aplicaciones actuales de la ciencia de datos y que ha cobrado gran relevancia, es el aprendizaje y posterior predicción del comportamiento de la demanda de productos, que puede estudiarse a partir de diferentes metodologías estadísticas enfocadas en el tiempo. Sin embargo, debido a la variabilidad de la demanda y de las variables que la afectan, es cada vez más complejo en el día de hoy realizar una proyección similar a la realidad, sobre todo si existe influencia por otros productos de la compañía o de la competencia. Ejemplo de esto puede ser la activación o creación de ofertas (promociones) que pueden producir como efecto una disminución en la demanda de los productos de línea, este fenómeno se conoce como "canibalización". El comportamiento de este fenómeno repercute gravemente a la planeación eficiente de la cadena de suministro, en cuyo caso, su incorrecto analisis y pronostico afecta a indicadores importantes como desguace (producto terminado vencido) y nivel de servicio.

Por lo anterior, fue necesario construir una solución asertiva para mejorar el entendimiento del fenómeno y para el desarrollo de una solución basada en un modelo de predicción. En primer lugar, se identificaron las variables que afectan la demanda de los productos de línea cuando existen promociones; Luego, se obtuvo la información pertinente de las variables con sus históricos, posteriormente se organizó y se estructuro adecuadamente esta información, de acuerdo a los lineamientos requeridos para la correcta predicción con el fin de aumentar la precisión del modelo a la realidad y Por último, se implementó el modelo y se realizaron pruebas de verificación y validación de los resultados para evaluar el rendimiento del modelo con la metodología actual. Todo lo anterior apoyado del equipo comercial y de monitoreo de la cadena de suministro de la compañía.

Los resultados del proyecto fueron aceptados por los expertos del negocio y se tomó la decisión de continuar perfeccionando la solución.

INTRODUCCIÓN

Debido al aumento de la información al interior y exterior de las organizaciones en los últimos años, las empresas están en constante esfuerzo de aprovechar esta para su mejoramiento en la toma de decisiones en los niveles estratégicos, tácticos y operativos dentro de las compañías.

Una de las situaciones problemáticas que tiene que afrontar y aprender una organización que comercializa bienes como lo son los alimentos, es el fenómeno conocido en mercadeo como “canibalización”. Este fenómeno como se profundizará posteriormente, afecta a las organizaciones en la medida de que la oferta o promoción de uno de sus productos no solo afecta a su competencia sino también a ella misma, ya que a pesar del aumento de la demanda del producto de oferta, consecuentemente se genera la disminución de uno o más productos de línea de su portafolio. Indicadores importantes como el desguace (producto terminado vencido) y nivel de servicio depende en gran proporción de la correcta gestión de este fenómeno.

En consecuencia, las empresas tienen la necesidad de desarrollar nuevas herramientas que les permita conocer el comportamiento de la nueva demanda y cuál es la relación entre los productos de oferta y los de línea.

De acuerdo con lo anterior, el presente trabajo se desarrolló con el objetivo de proponer un modelo de predicción de ventas, que permite estimar de la manera más asertiva el comportamiento de las ventas del producto de línea cuando se tiene la presencia de una oferta, tomando como punto de partida, la realidad y contexto de una empresa manufacturera de alimentos cárnicos de Medellín.

El alcance del proyecto se definió de acuerdo a los recursos de información, los recursos humanos y el tiempo de desarrollo y es el siguiente: Solución analítica basado en un modelo de predicción utilizando Machine Learning para un solo producto de oferta que afecta a un solo producto de línea en un perfil (cliente) especifico.

La metodología implementada fue el framework CRISP-DM, específico para la ciencia y minería de datos, que posteriormente se profundizará con más detalle.

Este informe busca un mejor entendimiento del concepto de “canibalizaciòn”, y como mejora la cadena de suministro, aumentando la eficiencia, la predicción y la toma de decisiones asertivas.

OBJETIVO GENERAL

Elaborar un modelo de predicción de ventas que permita estimar de forma más precisa el comportamiento de las ventas de productos de línea en la presencia de productos en oferta.

OBJETIVOS ESPECÍFICOS

- Identificar las variables que afectan la demanda de los artículos de línea cuando existen ofertas o promociones.
- Obtener, analizar y clasificar la información existente de las variables definidas.
- Evaluar modelos para la predicción de ventas de línea.
- Validar el resultado del modelo con el equipo comercial y monitoreo de la cadena de suministro.

MARCO TEORICO

En la presente sección se expondrán conceptos, antecedentes y avances importantes referentes al problema en aras de fundamentar las bases teóricas para el desarrollo exitoso del estudio, que servirán además como insumo para la definición metodológica y la interpretación de resultados.

Inteligencia Artificial

La inteligencia artificial (IA) no es para nada nueva, de hecho, ha sido desarrollada desde los 50’s, termino introducido por John McCarthy, en principio, enfocada en resolver problemas intelectuales de alta dificultad (Turing, 2009; McCarthy, 1990). Hoy en día, gracias a los desarrollos tecnológicos de las últimas décadas como la computación, el internet y el “Big Data”, se ha convertido como protagonista en resolver problemas de reconocimiento de patrones, aplicada en campos diversos como la radiología (Singh, 2018; Obermeyer et al., 2016), la investigación del cáncer (Cruz et al., 2006), el sector público (Sousa et al., 2019), entre otros.

La inteligencia artificial, funciona sin intervención humana, aprende e identifica patrones en los datos, logrando concluir a partir de ellos (Čerka et al., 2017). El desarrollo de los métodos más avanzados de la IA en los últimos años es resultado principalmente de la cantidad de datos que se crean diariamente más que en los algoritmos en si (Tecuci, 2012). A pesar de los avances en las diferentes ramas y herramientas de la IA, aún resulta incompleta su aplicación en la sociedad en general, esto, debido seguramente a que las nuevas tecnologías toman varios años en adaptarse (Brynjolfsson & Mitchell, 2017).

Machine Learning (ML)

Machine Learning es una rama de la inteligencia artificial que emplea varias herramientas estadísticas, probabilísticas y de optimización para aprender de los datos con la finalidad de clasificar nueva información, identificar patrones y nuevas tendencias. Es una herramienta poderosa porque, a diferencia de los métodos convencionales, permite realizar inferencias y decisiones que no se podría identificar y analizar de otra manera (Mitchell, 1997; Duda et al., 2001).

El éxito en el machine learning no está garantizado, este depende, de entender a la perfección el problema, junto a una exhaustiva apreciación y conociendo las limitaciones de los datos (Cruz et al., 2006).

Existen 3 diferentes tipos de ML: 1. Aprendizaje supervisado; 2. Aprendizaje no supervisado y 3. Aprendizaje reforzado. Clasificados en base a las “salidas” (outputs) que se deseen obtener (Mitchell, 1997; Duda et al., 2001). En el aprendizaje supervisado un proveedor (profesor) le entrega al algoritmo un conjunto de datos etiquetados, este conjunto de datos servirá como entrenamiento, donde el algoritmo intentará aprender de los datos para mapear la salida deseada. Por otro lado, en el aprendizaje no supervisado, se entrega al algoritmo un conjunto de datos no etiquetados, y en este caso, el algoritmo intentará encontrar patrones o descubrir grupos (Cruz et al., 2006). Esta categoría de aprendizaje automático se denomina no supervisado porque carece de una variable de respuesta que pueda supervisar el análisis (James et al., 2013)

El primer paso para determinar si se utiliza un método de machine learning y cual, en específico, depende de la pregunta de investigación (Jiang et al., s.f). Existen tres tareas

principales en la ciencia de datos: descripción, predicción e inferencia causal (Hernán et al., 2019).

Canibalización

En 1976, James Heskett definió la canibalización como el proceso por el cual un nuevo producto gana una porción de las ventas de otro producto existente (Heskett, 1976). Es decir, canibalización es definida como la reducción en el volumen de las ventas o cuota de mercado de un producto como el resultado de la introducción de un nuevo producto por el mismo productor (Kotler & Keller, 2012).

Hoy en día es elemento clave en el contexto comercial, principalmente en el análisis de la eficiencia de promociones u ofertas.

La teoría detrás de la canibalización recae en la teoría de consumo y de bienes sustitutos. En la teoría de consumo, dos bienes son sustitutos si un aumento en el precio del producto A causa un aumento en la demanda del producto B, ejemplos de lo anterior son: la mantequilla y la margarina, Coca-Cola y Pepsi, conos de helado y paletas. Esta posibilidad de sustitución resulta en canibalización promocional (Herrala, 2018).

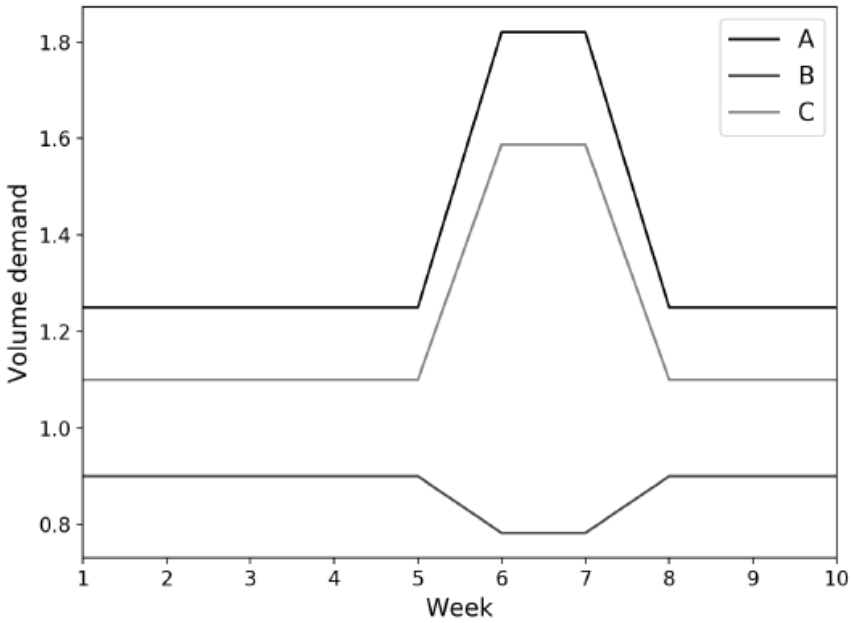

Ilustración 1: Serie de tiempo de ventas de los productos A, B y C en semanas (Herrala, 2018)

Como se puede observar en la ***Ilustración 1***, la canibalización resulta de las preferencias temporales de los consumidores en adquirir un producto por su precio, visibilidad o características, causando una disminución del volumen de ventas del producto existente “C”.

En resumen, cualquier producto nuevo deberá tomar una cuota de mercado de los demás jugadores (Ehrenberg, 1991) y predecir los efectos de esta canibalización es una tarea crítica y difícil. Ignorar los efectos de la canibalización puede dar como resultado graves consecuencias en el rendimiento financiero de una compañía (Harvey and Kerin, 1979; Chen and Yu, 2001).

Unos de los efectos más importantes de la canibalización de productos según Green y Krieger (1992) demuestran que la canibalización reduce la cuota de mercado de los productos canibalizados mientras que en general incrementa toda la cuota de mercado de la compañía.

Por lo anterior, el problema de medir la canibalización ha sido un gran foco de investigación. Lomax (1996) y Lomas et al. (1997) Identifican y miden la canibalización a través pruebas de análisis de ganancias / perdidas, duplicación de tablas de compra y las desviaciones de los movimientos de acciones esperados en los datos del panel de consumidores relacionados con las extensiones de línea en los mercados de detergentes del Reino Unido y Alemania, y sugieren la necesidad de que los gerentes utilicen múltiples métodos al evaluar el grado de canibalización.

METODOLOGÍA

Para el desarrollo de la propuesta de modelo de predicción, se utilizará la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), el cual es un modelo estándar abierto propuesto por IBM en 1999 para proyectos relacionados con minería de datos (IDECA, 2019). El marco metodológico CRISP-DM es ampliamente utilizado en la industria y a nivel empresarial por equipos de analítica como se puede observar en la ***Ilustración 2*** publicada por kdnuggets.com (Galán, 2015)

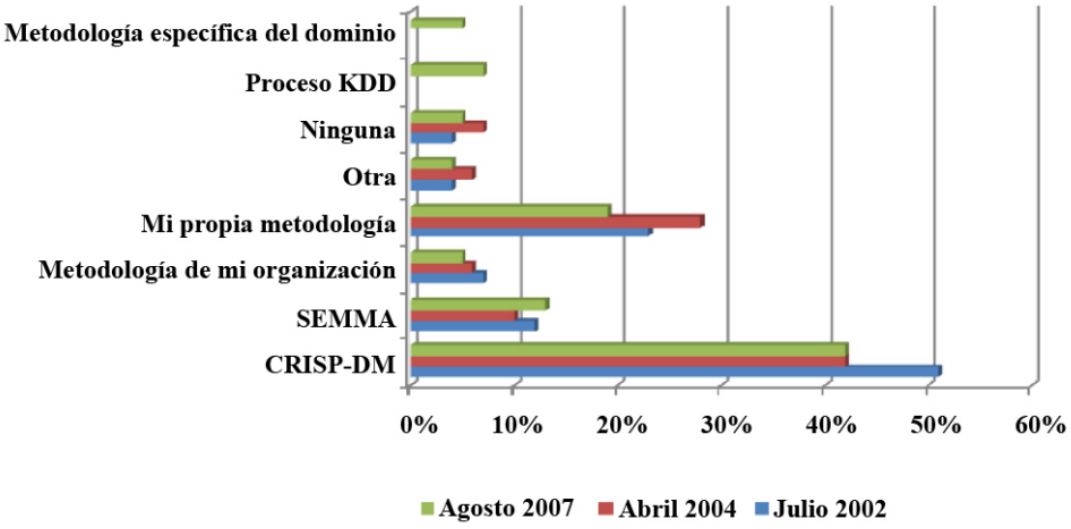

Ilustración 2: Uso metodologías para minería de datos (Elaborado por kdnuggets.com)

Hasta hoy en día, el modelo CRISP-DM se sigue utilizando y está basada en un modelo de proceso jerárquico que consta de seis fases, algunas de las fases son bidireccionales, lo que significa que es posible devolverse en el proceso para mejorar o resolver inconvenientes, por lo cual la sucesión de fases no tiene que ser ordenada. En la ***Ilustración 3*** se puede observar las fases de la metodología y las posibles secuencias entre ellas.

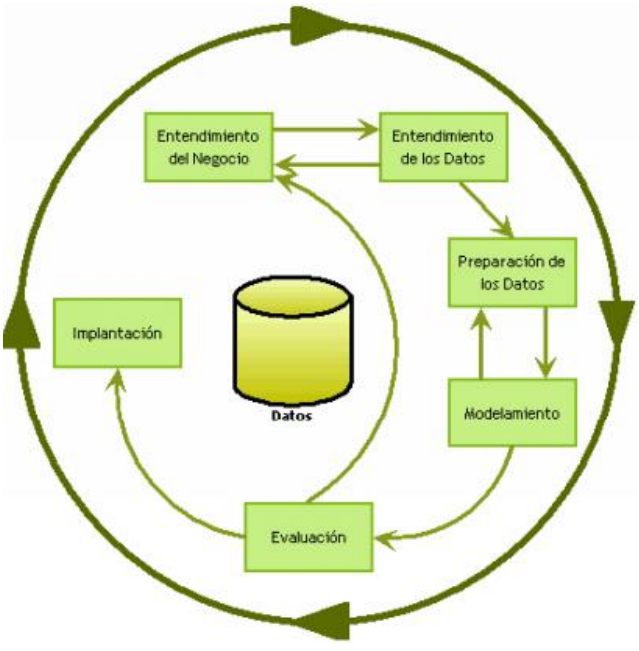

Ilustración 3: Modelo CRISP-DM (Elaborado por Galván, 2015)

A continuación, se explican cada una de las fases (Rodríguez, 2010):

1. Entendimiento del negocio: La primera fase es la más importante, ya que en esta se definen los objetivos del proyecto y requerimientos desde la perspectiva del negocio y es necesario convertir o transformar el problema o necesidad en un caso de uso de minería de datos (Wirth & Hipp, 2000).

2. Entendimiento de los datos: En esta etapa es donde se realiza un primer contacto con los datos, es el momento de familiarizarse con ellos, identificar la naturaleza de las variables, sus características y calidad de estas, como también establecer las relaciones entre ellas, esta fase junto a la siguiente es posiblemente la que requiera mayor tiempo.

   Esta fase está fuertemente relacionada con la fase anterior del entendimiento del negocio por lo cual es recomendable que las primeras hipótesis elaboradas a partir de los datos sean validadas con el entendimiento del negocio y los objetivos del proyecto (Wirth & Hipp, 2000).

3. Preparación de los datos: En esta fase, se desarrollan técnicas de limpieza, generación de variables, integración con otras fuentes de datos, cambios de formato, organización y preparación en general con el fin de tener un conjunto de datos de calidad para realizar una correcta modelación. Esta fase se relaciona con la siguiente "Modelamiento", ya que de acuerdo con la técnica que se desee implementar muy probablemente es necesario realizar cambios en los datos. (Wirth & Hipp, 2000).
4. Modelamiento: En esta fase, se seleccionan varias técnicas de minería de datos que constan de modelos estadísticos en su mayoría, la selección será principalmente basada en el caso de uso, como también en el problema que se desea resolver, la naturaleza y tipología de las variables y los datos, además del conocimiento y tiempo que requiere el analista para desarrollar el modelo. (Wirth & Hipp, 2000).

5. Evaluación: En esta fase se evalúa el modelo desde diferentes perspectivas, desde la perspectiva de la minería de datos en las cuales se revisan medidas de rendimiento y error, pero más importante aún desde la perspectiva del negocio, donde se valida los resultados del modelo con la realidad. En general se revisa todo el proceso anterior (Wirth & Hipp, 2000).
6. Implantación: En esta fase, una vez construidos los modelos, seleccionado el mejor y validado, se transforma el conocimiento obtenido en acciones de valor dentro del proceso de negocio, esto se logra a partir de la transformación del modelo en una infografía, visualización o hasta una aplicación que permita al usuario que requiera los resultados del modelo explotar o utilizar el modelo de forma adecuada y asertiva para la mejora en la toma de decisiones (Wirth & Hipp, 2000).

De acuerdo con la metodología anterior, se propone el siguiente desarrollo del informe para cumplir los objetivos:

1. Entendimiento del negocio.
2. Entendimiento de los datos.
3. Preparación de los datos
4. Modelamiento.
5. Evaluación.
6. Implantación.

RESULTADOS Y ANÁLISIS

1. Entendimiento del negocio.

El negocio lanza ofertas (promociones) al mercado y el equipo comercial define las cantidades, el tiempo en el cual se hará la activación y el valor de la canibalización en términos porcentuales. Cuando se tienen datos históricos, el porcentaje de canibalización se obtiene a través de ella, cuando no, la canibalización se da de forma intuitiva, es decir, desde el juicio y la experiencia humana. La necesidad del negocio radica en determinar si la canibalización es muy alta con respecto a lo que ocurre realmente y si se genera sobre producción, altos valores de desguace, se incurre en gastos logísticos, entre otros. Por otro

lado, si la canibalización es muy baja con respecto al comportamiento real del mercado, esto afecta: el nivel de servicio (muy bajo) y un costo elevado de oportunidad por no vender.

De acuerdo con lo anterior, el negocio y el departamento de planeación de la cadena de suministro se han fijado el objetivo de implementar un proyecto para aprovechar las nuevas tecnologías como la inteligencia artificial para obtener información significativa y valiosa de los datos que permitan estudiar el fenómeno de la canibalización de manera más asertiva y así mejorar los procesos e indicadores cómo el aumento en nivel de servicio y disminución en desguace.

Teniendo en cuenta la dimensión de la necesidad a resolver y el tiempo estimado en el desarrollo del proyecto, el equipo de demanda y de monitoreo, se fijan un alcance de un primer piloto de baja escala a partir de una sola referencia de oferta y un solo perfil (cliente).

1.1 Identificación de variables

En primera instancia, para la identificación y definición de variables, se tomó en consideración variables descritas en diversos artículos académicos como también variables que junto a un equipo interdisciplinar de la compañía se definieron como variables importantes a considerar por su alta probabilidad de impacto de la canibalización. De acuerdo con lo anterior y teniendo en cuenta la información histórica disponible de ventas y oferta, se realizaron los ajustes pertinentes para organizar la información.

A continuación, se presenta una tabla con las variables identificadas, las cuales fueron evaluadas y de acuerdo con su disponibilidad, entendimiento y requerimientos serán seleccionadas o no como insumo para el ejercicio:

Variable	**Descripción**
Referencia de línea	Es la referencia de línea impactada, es decir, la que sufre la canibalización
Tipo de oferta	Amarre, extra-contenido o combo virtual
Cliente	Es la organización (almacén, mayorista, minorista, etc.) al cual se le vende la oferta y la referencia de línea

CEDI	Es el centro de distribución desde donde se efectúa la venta
Marca	Es la marca de la referencia ofertada
Mes	Mes en el cual inicio el evento
Duración	Duración del evento en semanas
Tipo de evento	Se especifica si el evento es triple A (eventos publicados en boletín generalmente en aniversarios de los clientes) o si se trata de un evento convencional (activados en su mayoría para incrementar flujo en los puntos de venta, o por estrategia de bloqueo a la competencia).
Descuento	Es el descuento aplicado a la oferta
Cod_Canal	Canal de ventas
Cod_SubCanal	SubCanal de ventas
Canibalización	Son las unidades dejadas de vender de la referencia de línea por cada unidad vendida de la oferta.

Tomando como punto de partida la definición de las variables anteriores, se da paso al entendimiento de los datos, identificando particularidades de cada variable, para posteriormente procesarlos y obtener un conjunto valioso de información.

2. Entendimiento de los datos

De acuerdo con el entendimiento de la necesidad y a la disponibilidad de la información, se obtuvo la siguiente información de bases de datos de la compañía como SQL Server principalmente.

Tabla 1: Información (Elaboración propia)

Año natural/Semana	centro	perfil	VentaLinea	VentaOferta
2018.01	NN13	AUT INDEPENDIENTE	487.0	240.0
2018.01	NN24	PUNTO DE VENTA	49.0	NaN
2018.01	NN24	TRAD CCIALIZADORES	150.0	NaN
2018.01	NN24	TRAD DIRECTO	1.0	NaN
2018.01	NN27	AUT INDEPENDIENTE	2270.0	NaN
...	...	...	...	...
2020.01	NN18	TRAD CCIALIZADORES	160.0	NaN
2020.01	NN18	SURTIMAX	211.0	NaN
2020.01	NN18	OLIMPICA	301.0	NaN
2020.01	NN24	MAKRO-ALKOSTO	50.0	NaN
2020.01	NNB8	OLIMPICA	20.0	NaN

Como se puede identificar en la ***Tabla 1***, tenemos el historial de ventas de la referencia de oferta y las ventas de la referencia de línea correspondiente por semana de todos los centros de distribución a todos los perfiles (clientes) por un periodo de dos años 2018-2019.

Tabla 2: Métricas descriptivas de variables cuantitativas (Elaboración propia)

	VentaLinea	VentaOferta
count	11319.000000	1168.000000
mean	597.544129	772.585616
std	1123.303842	1230.647189
min	0.000000	1.000000
25%	42.000000	90.000000
50%	160.000000	330.000000
75%	590.000000	854.250000
max	15092.000000	12136.000000

De acuerdo con la ***Tabla 2***, se puede identificar alta variabilidad en las variables cuantitativas, que nos interesan, las cuales son VentaLinea y VentaOferta de todos los centros y perfiles, con coeficientes de variación de 1,88 y 1,59 respectivamente. Lo anterior, era lo esperado, ya que se cuenta con centros de distribución principales y perfiles más grandes que otros.

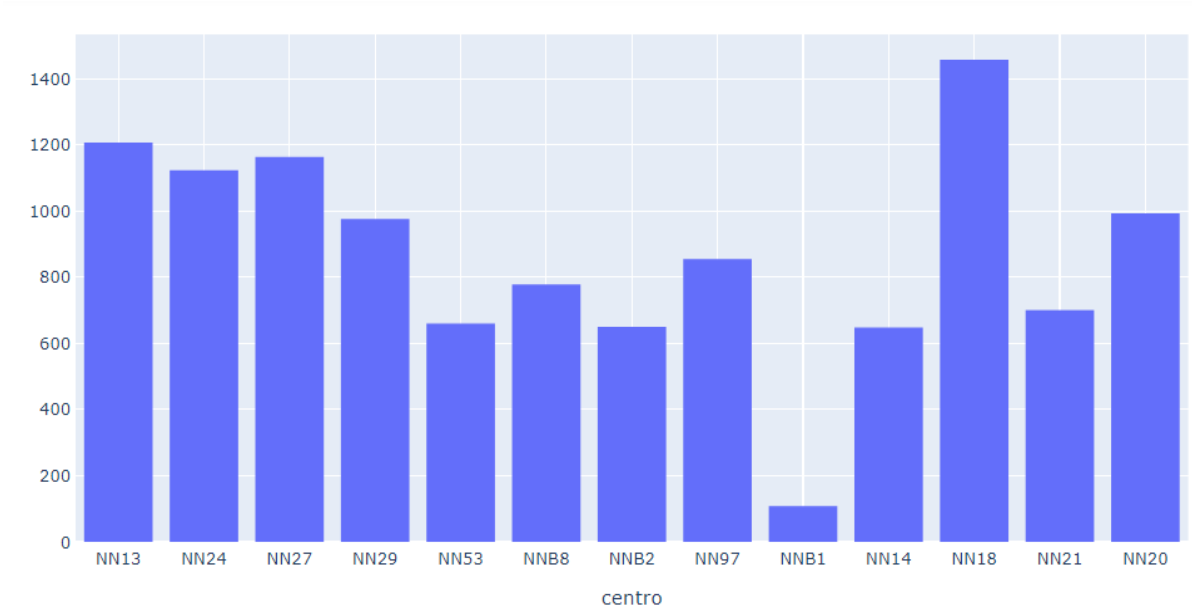

Ilustración 4: Histograma por centro (Elaboración propia)

De acuerdo con la ***Ilustración 4*** y el número de pedidos, se puede afirmar que la mayor rotación del par línea-oferta se da de los centros principales como "NN18" (Bogotá) y "NN13" (Barranquilla) y el de menor rotación es "NNB1".

Otro hallazgo importante son los perfiles con mayor cantidad de demanda de estos productos, se destacan el canal "AUT INDEPENDIENTE", "EXITO" y "OLIMPICA" como se puede observar en la ***Ilustración 5*** como eje y el número de pedidos.

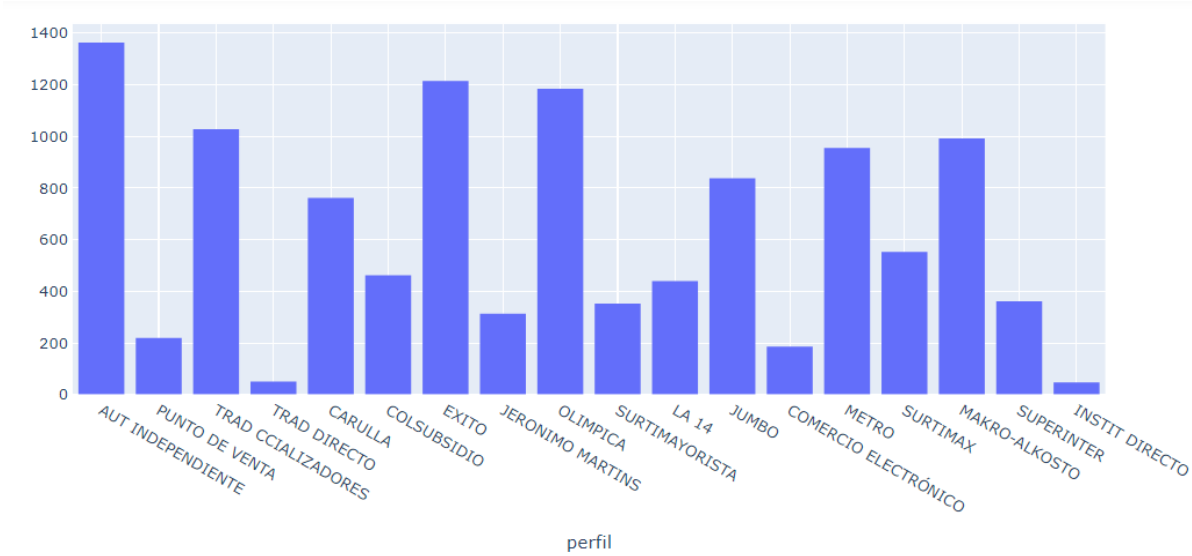

Ilustración 5: Histograma por perfil (Elaboración propia)

3. Preparación de los datos

Los equipos de monitoreo y de demanda, definieron que la variabilidad de la venta de oferta por semana es muy alta debido al comportamiento del mercado y a la ejecución de actividades externas del perfil, por lo cual, se tomó la decisión de realizar agregación por evento, definiendo el evento, como el conjunto de semanas consecutivas donde existió venta oferta en un perfil.

El perfil seleccionado para realizar un primer modelo exploratorio fue el perfil "EXITO", definido de igual forma por los expertos del equipo de demanda, este perfil fue seleccionado por que tiene una gran cantidad de eventos con respecto a otros perfiles en estas referencias particularmente, además de tener en cuenta el alcance real del proyecto y la disponibilidad de la información.

Realizando los procedimientos anteriormente descritos, obtenemos un conjunto de datos de 90 registros (eventos), como se puede observar en la ***Tabla 3***.

Tabla 3: Datos por evento (Elaboración propia)

centro	perfil	mes_inicio_evento	duracion_evento	Vta_linea	Vta_oferta	linea/oferta	vta_linea/semana	vta_oferta/semana
NN13	EXITO	1	1	284	456	0.622807	284	456
NN13	EXITO	5	2	6	1746	0.003436	3	873
NN13	EXITO	8	3	264	1406	0.187767	88	468
NN13	EXITO	11	2	1664	2367	0.703000	832	1183
NN13	EXITO	4	1	916	1000	0.916000	916	1000
...	...	...	...	...	...	...	...	...
NNB8	EXITO	4	1	663	289	2.294118	663	289
NNB8	EXITO	4	1	82	162	0.506173	82	162
NNB8	EXITO	7	1	383	115	3.330435	383	115
NNB8	EXITO	8	1	587	128	4.585938	587	128
NNB8	EXITO	12	1	436	108	4.037037	436	108

En el transcurso de los "sprint's" entre la etapa de modelación y preparación de los datos, se obtuvo como conclusión la necesidad de "dumificar" o discretizar la variable categórica "centro", esta técnica se explicará a continuación:

- 'Dumificar': Esta transformación de los datos, es realizada en algunos casos, para la etapa de modelación, esta técnica consiste en transformar una variable categórica como lo es la

variable “centro”, en varias variables cuantitativas, como se podrá identificar en la ***Ilustración 6.***

centro	venta
NN13	1000
NN18	2000
NN21	1000
NN13	1500

venta	centro NN13	centro NN18	centro NN21
1000	1	0	0
2000	0	1	0
1000	0	0	1
1500	1	0	0

Ilustración 6: Tablas dumificadas (Elaboración propia)

De acuerdo con la ilustración anterior obtenemos de la variable categórica “centro”, 12 variables cuantitativas binarias como se observa en la ***Tabla 4***.

Tabla 4: Dataset final listo para Modelación (Elaboración propia)

vta_oferta/semana	Vta_oferta	linea/oferta	centro_NN13	centro_NN14	centro_NN18	centro_NN20	centro_NN21	centro_NN24	centro_NN27	...
456	456	0.622807	1	0	0	0	0	0	0	...
873	1746	0.003436	1	0	0	0	0	0	0	...
468	1406	0.187767	1	0	0	0	0	0	0	...
1183	2367	0.703000	1	0	0	0	0	0	0	...
1000	1000	0.916000	1	0	0	0	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...
289	289	2.294118	0	0	0	0	0	0	0	...
162	162	0.506173	0	0	0	0	0	0	0	...
115	115	3.330435	0	0	0	0	0	0	0	...
128	128	4.585938	0	0	0	0	0	0	0	...
108	108	4.037037	0	0	0	0	0	0	0	...

4. Modelación

Una vez obtenido el DataSet final preparado, se procede a realizar la técnica para modelación “train-test split” o “separación de entrenamiento y prueba”, este consiste en un procedimiento que se efectúa para estimar el rendimiento de un modelo. Básicamente, es la división o partición del conjunto de datos en dos partes, una de ellas para entrenar el modelo (70-80% del DataSet) y la otra de ellas para pruebas (20-30%). De acuerdo con lo anterior, obtenemos un conjunto de entrenamiento (Training Set) de 72 datos (80%) y un conjunto de prueba (Test Set) de 18 datos (20%).

Número total de datos (100%)	
Training Set (80%)	Test Set (20%)

Training Set (80%)			
X_train			y_train
X1	X2	Xn	y

Test Set (20%)				
X_test			y_test	y_pred
X1	X2	Xn	y	y'

Ilustración 7: Partición del DataSet (Elaboración propia)

Como se puede observar en la ***Ilustración 7***, tanto el Training Set como el Test Set se divide en X_train, y_train, X_test y y_test, respectivamente, el modelo aprenderá del Training Test y realizará la prueba prediciendo de acuerdo a X_test, dando como resultado y_pred, la cual se compara con y_test, obteniendo el rendimiento del modelo y su ajuste.

Se realizo una extensa y completa investigación y experimentación en herramientas de inteligencia artificial como Machine learning y Deep learning para la modelación del conjunto de datos, por simplificación se describirá el modelo con mejor resultado global.

El tipo de modelación que se le realizaron a los datos fue Deep learning con redes neuronales artificiales, esta modelación se explicara a continuación:

Las redes neuronales artificiales son un conjunto de modelos simplificados que emulan el comportamiento de manera abstracta de cómo funciona el cerebro humano (Salas, s.f).Como se puede observar en la ***Ilustración 8***, las neuronas reciben información por medio de entradas, estas interpretan la información y dan una salida o resultado a otra neurona.

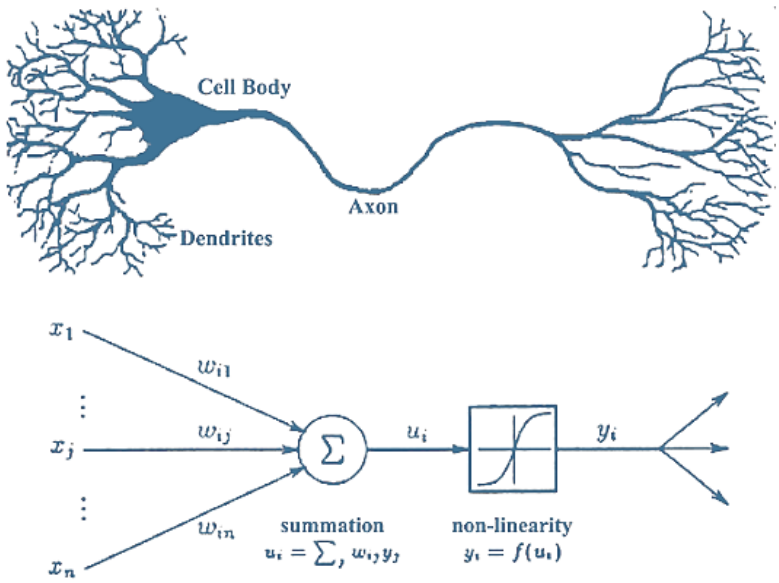

Ilustración 8: Red neuronal artificial. Por: Magiquo (2019).

Las redes neuronales artificiales se componen en tres partes, la capa de entrada, una o más capas ocultas y una capa de salida. (Florez, R., & Fernandez, J., 2008).

El procedimiento que las redes neuronales artificiales realizan con los datos se pueden separar en dos momentos, forward propagation (propagación hacia adelante) en la cual se realiza una ponderación de las entradas con pesos aleatorios y se realiza una función de activación y así entre cada neurona en cada capa de la red y un segundo momento denominada back propagation (propagación hacia atrás), en el cual se optimiza la función de costo y pesos, minimizando el error y la pérdida de valor (Basogain, X., s.f).

El modelo que se utilizó en este proyecto en particular fue el perceptrón regresivo multicapa (MLPRegressor), el cual es una de las redes más simples y usadas.

Para el ejercicio se optimizo la red en cuatro capas ocultas de tamaños de neuronas 8, 256, 128 y 16 respectivamente, con un optimizador "Adam" que combina el AdaGrad, el cual es un algoritmo de gradiente adaptativo con el RMSProp propagación de raíz cuadrada media que remplazan el descenso de gradiente estocástico, Adam es más rápido y ampliamente utilizado; el número de iteraciones final fue de 1000 con tamaño de lote de 4 y una tasa de aprendizaje de 0.01.

5. Evaluación

Desde la perspectiva de la minería de datos, se evaluó la métrica de rendimiento R2 o coeficiente de determinación para el conjunto de prueba, el cual fue de 0.7856, un resultado sobresaliente con respecto a otras modelaciones. A continuación, en la ***Tabla 5***, se podrá identificar la predicción de "y" realizada con el modelo y la prueba real de "y".

Tabla 5: Comparación de valores predictivos contra valores reales (Elaboración propia)

y_pred	y_test
1.87	1.15
1.64	3.01
1.38	2.34
5.40	4.73
3.66	3.40
0.46	0.91
4.16	3.11
2.48	1.50
1.41	1.75
4.03	3.8
3.43	2.59
1.62	2.42
1.50	1.10
4.8	4.23
1.4	1.30
0.29	0.73
6.48	4.35
6.78	4.89

Los equipos de demanda y de monitoreo evaluaron el rendimiento y la presicíon predictiva del modelo e identificaron la gran utilidad del modelo, que permite de manera mas cercana a la realidad entender el comportamiento de la canibalización de las referencias oferta y de linea especifica para el perfil ÉXITO..

6. Implementación

Se establecieron relaciones en el transcurso de un periodo de 6 meses con la unidad de analítica avanzada de una filial del grupo empresarial que presta servicios de tecnología, con ella se crearon acuerdos para continuar la industrialización e implementación de esta solución a escala para su posterior uso real. Por lo cual, para el alcance definido en este informe, no se abarco a cabalidad esta fase del proceso.

CONCLUSIONES

- Las fases de entendimiento del negocio y de los datos, permitieron identificar las variables más importantes a tener en cuenta en el estudio de la canibalización de productos, sin embargo, por la complejidad en el acceso a la información y en algunos casos escasez de la misma, se extrajo un subconjunto valioso de estas variables para desarrollar en este informe un “proxi” de predicción.

- Se logro aplicar técnicas de exploración de datos para analizar y clasificar las variables mas significativas para el entendimiento del problema, obteniendo ideas y conclusiones importantes a tener en cuenta en el desarrollo de la transformación y modelado de datos.
- Se evaluaron diferentes técnicas de Machine Learning, con el fin de obtener ideas y resultados de valor acerca del fenómeno de canibalizaciones, cómo se evidencio en el presente informe, por simplicidad se observa el desarrollo de la modelación con mejor resultado general y global.
- Los resultados de este proyecto, se compartieron con los equipos de monitoreo y de demanda de la cadena de suministro de la compañía y se aprobó la utilidad del modelo y se definió una hoja de ruta con la unidad de analítica avanzada para hacer pruebas especificas de validación y posterior implantación o industrialización para usuario final como meta para el primer semestre del 2021.

REFERENCIAS BIBLIOGRAFICAS

- Turing, Computing machinery and intelligence amplification. “parsing the turing test.”, Comput Intell Expert Speak Springer (2009) 25–44, https://doi.org/10.1109/9780470544297.ch3.
- John McCarthy, Artificial intelligence, logic and formalizing common sense, Philos Log Artif Intell (1990) 161–190.
- G. Singh, S.J. Al’Aref, M. van Assen, et al., Machine learning in cardiac CT: basic concepts and contemporary data, JCCT (2018).
- Z. Obermeyer, E.J. Emanuel, Predicting the future — big data, machine learning, and clinical medicine, N. Engl. J. Med. 375 (2016) 1216–1219, https://doi.org/10.1056/NEJMp1606181.Predicting.
- Sousa, W. G. de, Melo, E. R. P. de, Bermejo, P. H. D. S., Farias, R. A. S., & Gomes, A. O. (2019). How and where is artificial intelligence in the public sector going? A literature review and research agenda. Government Information Quarterly, 36(4), 101392. https://doi.org/10.1016/j.giq.2019.07.004

- Čerka, P., Grigienė, J., & Sirbikytė, G. (2017). Is it possible to grant legal personality to artificial intelligence software systems? Computer Law and Security Review, 33(5), 685–699. https://doi.org/10.1016/j.clsr.2017.03.022.
- Tecuci, G. (2012). Artificial intelligence. Wiley Interdisciplinary Reviews: Computational Statistics, 4(2), 168–180. https://doi.org/10.1002/wics.200.
- Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. Science, 358(6370), 1530–1534. https://doi.org/10.1126/science.aap8062.
- Cruz, J. A., & Wishart, D. S. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics, 2, 117693510600200. https://doi.org/10.1177/117693510600200030
- *Mitchell T. 1997. Machine Learning. New York: McGraw Hill.
- Duda RO, Hart PE, Stork DG. (2001) Pattern classification (2nd edition). New York: Wiley.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer-Verlag www.springer.com/us/book/9781461471370
- T. Jiang, J. L. Gradus and A. J. Rosellini, Supervised Machine Learning: A Brief Primer, Behavior Therapy, https://doi.org/10.1016/j.beth.2020.05.002
- Hernán, M. A., Hsu, J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. Chance, 32(1), 42–49. https://doi.org/10.1080/09332480.2019.1579578
- James L. Heskett. Marketing. Macmillan, (1976). ISBN 978-0-02-353940-4. Google-Books-ID: sO0TAQAAMAAJ
- Kotler, P., Keller, K., 2012. Marketing Management. Prentice Hall.
- Kwiatkowski, D., Phillips, P.C., Schmidt, P., Shin, Y., (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? J. Econom. 54 (1), 159–178.
- Herrala, O. (2018). Evaluating cannibalization between items in retail promotions (Bachelor's thesis: Engineering Physics and Mathematics). Aalto Universituy, Finland. https://sal.aalto.fi/publications/pdf-files/ther18_public.pdf

- Green, P. E., & Krieger, A. M. (1992). An application of a product positioning model to pharmaceutical products. Marketing Science, 11(2), 117–132.
- Lomax, W. (1996) The measurement of cannibalisation. Marketing Intelligence and Planning 14(7), pp. 20-28
- Lomax, W., K. Hammond, R. East, M. Clemente (1997) The measurement of cannibalisation. Journal of Product and Brand Management 6(1), pp.27-39
- IDECA. (2019, abril). Metodología para la Analítica de datos. ideca.gov.co. https://ideca.gov.co/sites/default/files/MetodologiaAnaliticaDatos.pdf
- Galán, V. (2015, octubre). APLICACIÓN DE LA METODOLOGÍA CRISP-DM A UN PROYECTO DE MINERÍA DE DATOS EN EL ENTORNO UNIVERSITARIO. Universidad Carlos III de Madrid - Escuela Politécnica Superior - Ingeniería en Informática. https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf
- Rodríguez, O. (2010). Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM. oldemarrodriguez. http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037
- Wirth, R. Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf
- Salas, R. (s.f). Redes Neuronales Artificiales. Departmento de computación, Universidad de Valparaíso. https://bit.ly/3qUFQck