Implementación de modelo de aprendizaje automático para el ordenamiento de pasillos en un supermercado digital basado en el contenido del carrito de compras del usuario

Daniela Ramos García, Gaymundo Guzmán Mata, Luis Antonio Barajas Ramirez, Sergio Ortiz Malpica, Arturo Durán Castillo

Filiación

Tecnológico de Monterrey

Contexto

En la actualidad, los supermercados enfrentan el reto de mejorar la experiencia de compra de sus clientes mientras optimizan sus operaciones internas. Con el crecimiento del comercio electrónico y la evolución de los hábitos de consumo, se vuelve cada vez más importante entender cómo las personas compran, qué productos combinan frecuentemente y cómo se mueven dentro de una tienda. En este contexto, la minería de datos y el análisis de cestas de mercado han cobrado relevancia como herramientas para identificar patrones de compra y proponer mejoras en la organización de los productos. A través del uso de algoritmos avanzados, es posible descubrir combinaciones de productos con alta utilidad, incluso si no se venden con frecuencia, y utilizar esta información para re-diseñar el layout de las tiendas, reducir tiempos de recorrido y aumentar la satisfacción del cliente. Este proyecto busca explorar estas posibilidades mediante la aplicación de técnicas de minería de itemsets de alta utilidad y optimización dinámica del layout, con el objetivo de lograr una experiencia de compra más eficiente y personalizada.

Planteamiento formal/técnico del problema

La experiencia de compra en línea no solo abarca la interfaz de usuario, si no también implica mucho en conocer el cliente y su comportamiento a la hora de realizar una compra. En la actualidad, los supermercados enfrentan el desafío de mejorar la experiencia de compra de sus clientes mientras optimizan sus operaciones internas. Con el crecimiento del comercio electrónico y la evolución constante de los hábitos de consumo, el reto de optimizar este proceso de compra en el usuario se ha vuelto más grande. Esto exige un mayor entendimiento de los patrones de compra, de las combinaciones frecuentes de productos y el comportamiento del cliente dentro de la tienda.

Esta investigación se propone aplicar técnicas avanzadas, combinándolas con métodos de optimización dinámica del layout de tienda, con el fin de generar una propuesta que permita rediseñar la organización de los pasillos, dado un patrón de compra, de este modo reduciendo los tiempos de recorrido y aumentando la satisfacción del cliente.

La relevancia de este trabajo radica en que contribuye tanto a la mejora de la eficiencia operativa como a la personalización de la experiencia de compra, generando beneficios económicos mediante decisiones informadas basadas en datos de utilidad. Asimismo, la motivación del estudio se fundamenta en la posibilidad de adaptar y validar modelos ya existentes en nuevos entornos,

integrando enfoques recientes como algoritmos evolutivos, mecanismos de seguimiento de usuarios y modelos predictivos de comportamiento, lo cual ofrece una oportunidad única para innovar en la gestión de espacios comerciales físicos y digitales.

Estado del arte

A survey of incremental high-utility itemset mining (Gan et al., 2018)

Este artículo da una revisión de los algoritmos de minería incremental de itemsets de alta utilidad (iHUIM por sus siglas en inglés), incluyendo enfoques de estos basados en Apriori, árboles y listas de utilidad, y aborda los principales desafíos y cuestiones de investigación en esta área. El artículo llega a esto ya que menciona que la minería tradicional de reglas de asociación (esto es una técnica que busca patrones en lo que las personas compran juntas) no es adecuada para aplicaciones reales donde se deben considerar factores como las ganancias unitarias y las cantidades de compra. La minería de itemsets de alta utilidad (HUIM por sus siglas en inglés) busca encontrar patrones rentables considerando estos factores, pero también dice que muchos algoritmos están diseñados para bases de datos estáticas. En aplicaciones reales, como el análisis de cestas de mercado, las bases de datos se actualizan dinámicamente con nuevas transacciones, lo que hace que los algoritmos de minería incremental de itemsets de alta utilidad (iHUIM) sean más eficientes al actualizar los HUIs de manera incremental, reduciendo los costos. Este será un buen artículo a leer ya que toma un punto de vista diferente e innovador a lo que otros artículos relacionados a este tema mencionan, será un buen artículo para abordar la solución desde otro punto de vista.

Crowd Distance Induced Multi Objective Binary Salp Swarm Optimization Algorithm for Mining High Frequency and Utility Itemsets (Budaraju, R.R., 2025)

El estudio de este artículo trata sobre cómo encontrar combinaciones de productos importantes en bases de datos de ventas. Para eso, no solo se fijan en qué productos se venden juntos muchas veces, sino también en qué tan útiles o rentables son. Usan un nuevo algoritmo llamado CD-BSSOA, que ayuda a buscar esas combinaciones valiosas equilibrando bien entre probar nuevas opciones y aprovechar las buenas. Este método es útil en cosas como analizar compras de clientes, hacer mejores recomendaciones de productos y manejar inventarios. Al probarlo con datos reales, el algoritmo mostró mejores resultados que otros métodos anteriores, encontrando más soluciones útiles y con mejor rendimiento. Será un buen artículo para abordar la solución desde un punto de vista más específico y aprendiendo sobre una solución ya probada.

HURI - A novel algorithm for mining high utility rare itemsets (Pillai, J.,Vyas, O.P.,Muyeba, M., 2013)

El artículo aborda el algoritmo HURI para la Minería de Itemsets Raros de Alta Utilidad (HURI), que genera itemsets raros y de alta utilidad según los intereses del usuario, mediante un enfoque de dos fases. La minería de utilidad busca identificar itemsets valiosos, incluso aquellos con baja frecuencia de venta pero alta rentabilidad. El trabajo evalúa el rendimiento y la complejidad del algoritmo HURI, demostrando su eficiencia en la identificación de itemsets útiles para la toma de decisiones, especialmente en análisis de cestas de mercado.

Data-driven personalized assortment optimization by considering customers' value and their risk of churning: Case of online grocery shopping.

El artículo propone un modelo llamado CA&C (Customized Assortment considering Churn), que optimiza el surtido de productos en supermercados en línea tomando en cuenta tanto las preferencias de los clientes como su riesgo de abandono. El núcleo del trabajo está en priorizar el inventario limitado para clientes valiosos y en riesgo, mediante técnicas como modelos logit multinomiales, análisis de supervivencia y programación dinámica. Entre los principales resultados, se destaca que el modelo redujo la tasa de

abandono hasta en 11.7 % en comparación con políticas convencionales, y aumentó las ganancias entre un 8 % y un 24 %, con un promedio del 17 %. Además, se comprobó que la utilidad percibida por género y día de la semana influye en las decisiones de personalización, y que variables como la frecuencia de compra y el tiempo entre visitas (TBS) son claves para anticipar el abandono. Como principal debilidad, el modelo aún no ha sido validado completamente en entornos reales fuera del caso específico de una tienda de carne, lo que limita la generalización de sus beneficios.

Modeling the Optimal Grocery Store Trading Area Using Machine Learning Methods

Se enfoca en definir el tamaño ideal del área de ventas en supermercados físicos, utilizando técnicas como regresión lineal múltiple y clustering de series temporales. Su contribución principal es la identificación de un rango óptimo de densidad de clientes entre 1.5 y 2.2 metros cuadrados por persona en horas pico, lo que permite maximizar ingresos y mantener una experiencia de compra cómoda. El estudio evidencia que reducir el espacio por debajo de ese umbral puede disminuir las ventas hasta en un 12.8 %, mientras que aumentarlo más allá de los 3 m² por cliente no genera mejoras significativas. Esta propuesta es útil para el diseño eficiente de tiendas y la toma de decisiones en proyectos de inversión. Sin embargo, el trabajo tiene debilidades metodológicas: no se especifica el tamaño ni la diversidad del conjunto de datos utilizado, y no se consideran factores conductuales más complejos como motivaciones de compra o la integración con medios digitales.

Customer Behavior in an Online Ordering Application: A Decision Scoring Model

El artículo desarrolla un modelo de puntuación predictiva para anticipar la probabilidad de recompra de clientes en un entorno de supermercado en línea. Utilizando una combinación de encuestas subjetivas y datos reales de comportamiento de compra de los clientes a lo largo de 12 meses, el estudio analiza a 1,089 clientes de dos empresas americanas. A través de regresión logística multinomial, se identificó que las variables con mayor poder predictivo fueron la frescura del producto (PF), el ahorro de tiempo (TS) y el número de pedidos previos. La precisión del modelo fue del 59.2 %, superando ampliamente el azar (35.9 %). Aunque también se evaluaron la calidad del servicio (SQ), la calidad del producto (PQ), la facilidad de uso del sitio (SE) y la sensibilidad al precio, algunas de estas no fueron significativas en todos los contextos, especialmente el precio. Sin embargo, una limitación del articulo es que se basa en percepciones de los clientes, que pueden no reflejar su comportamiento real, y no considera factores externos como promociones o acciones de la competencia.

An E-Business Event Stream Mechanism for Improving User Tracing Processes

En este artículo propone un mecanismo innovador para mejorar el rastreo de usuarios en plataformas de comercio electrónico mediante el uso de flujos de eventos y fórmulas de lógica temporal lineal. El núcleo del trabajo radica en la creación de un “universo de eventos” estructurado con cinco parámetros clave (usuario, categoría, nivel, tiempo y acción), lo que permite registrar de forma precisa las actividades del usuario y predecir su comportamiento futuro. La metodología se valida aplicando algoritmos de clustering y clasificación sobre dos conjuntos de datos reales (Online Shoppers e Instacart), logrando altos niveles de precisión en la predicción de preferencias de usuario. Entre sus principales contribuciones destacan: la formalización de relaciones de satisfacción para modelar procesos de usuario, la integración de técnicas de minería de datos con lógica formal, y la validación empírica con métricas como precisión, recall y F1-score. No obstante, el enfoque presenta algunas debilidades, como la dependencia de estructuras de datos bien definidas y la falta de pruebas en entornos más dinámicos o con datos no estructurados, lo que podría limitar su aplicabilidad en escenarios reales más complejos.

Optimization of store layout using market basket analysis

En este artículo se propone una metodología basada en el Market Basket Analysis para optimizar la disposición de productos en tiendas minoristas, utilizando datos de ventas obtenidos de sistemas de punto de venta. El núcleo del trabajo radica en aplicar técnicas de minería de datos, específicamente el algoritmo Apriori, para identificar reglas de asociación entre productos que los clientes suelen comprar juntos. A partir de estas reglas, se sugieren estrategias para reorganizar el layout de la tienda, como ubicar productos de alto margen cerca de secciones de alta demanda, o agrupar productos complementarios en zonas promocionales.

Entre sus principales contribuciones destaca el uso de herramientas de visualización en R para representar gráficamente las asociaciones encontradas, lo que facilita la toma de decisiones para los minoristas. Además, el estudio demuestra cómo el análisis de 5147 transacciones puede revelar patrones de compra útiles para mejorar la experiencia del cliente y aumentar las ventas. Sin embargo, una debilidad importante es que el análisis se basa en datos de solo una semana, lo que limita la generalización de los resultados y no considera variaciones estacionales ni diferencias entre tipos de tiendas. Tampoco se profundiza en la validación práctica de los cambios propuestos en un entorno real de tienda.

Design of facility layout with lean service and market basket analysis method to simplification of service process in the supermarket

En este artículo se presenta una propuesta metodológica para mejorar la eficiencia del servicio en un supermercado mediante la integración de dos enfoques complementarios: Lean Service, que permite identificar y eliminar actividades que no agregan valor, y Market Basket Analysis, que analiza los patrones de compra de los clientes para optimizar la disposición de los productos. El estudio identifica siete tipos de desperdicio en el proceso, como tiempos de espera, errores en documentos y movimientos innecesarios, y logra reducir el tiempo total del proceso de 319.62 a 200.76 horas, aumentando la eficiencia del ciclo al 48.85%. A través del MBA, se detectaron 14 productos con relaciones significativas, lo que permitió diseñar tres alternativas de layout, seleccionando finalmente la opción que ubicaba productos de alta demanda al frente de la tienda. Entre sus principales contribuciones se encuentra la aplicación combinada de herramientas de ingeniería industrial para rediseñar procesos en el sector retail, mejorando tanto la experiencia del cliente como la eficiencia operativa. Sin embargo, una limitación importante es que el estudio se basa en datos de una sola empresa y un periodo específico, lo que restringe la generalización de los resultados a otros contextos o tipos de supermercados.

Objetivos de la investigación

General

Desarrollar y validar un modelo de aprendizaje automático que, a partir del contenido del carrito de compras de un usuario y el histórico de pedidos realizados en la aplicación, prediga los pasillos con mayor probabilidad de ser visitados, esto con el fin de organizar los pasillos dentro de cada departamento con el propósito de minimizar el esfuerzo de navegación del comprador y mejorar la eficiencia de llenado de carritos.

Específicos

1. Explorar el conjunto de datos: Analizar la estructura y calidad de los archivos proporcionados, además de identificar valores faltantes e inconsistencias que puedan afectar al modelado.
2. Definir la métrica de esfuerzo de navegación: Formalizar un algoritmo que calcule el esfuerzo basado en la posición vertical de cada pasillo y el número de desplazamientos necesarios para acceder a él.
3. Diseñar y entrenar el modelo de ranking de pasillos: Seleccionar los algoritmos adecuados a implementar.
4. Integrar el modelo para su validación: Desarrollar una simulación en la que se reordenen los pasillos en

tiempo real cada vez que el carrito se actualice para posteriormente medir el esfuerzo para realizar los pedidos con el nuevo ordenamiento.

5. Comparar el desempeño frente a otros acomodos: Cuantificar la reducción de esfuerzo promedio respecto a ordenamientos aleatorios con el layout propuesto. Aplicar pruebas para confirmar la significancia de las mejoras.

Pregunta(s) de investigación e Hipótesis.

- ¿Cuál es el promedio en cantidad de 'scrolls' de los usuarios para seleccionar un producto?
- ¿Qué departamentos tienen mayor tendencia en dado momento del día?
- ¿Qué modelo de machine learning es el más óptimo para reorganizar productos y recomendar cosas al usuario?
- ¿Cómo podemos medir el esfuerzo del usuario?
- Actualmente, ¿cómo se organizan los pasillos de los departamentos?
- ¿Cómo podemos correlacionar departamentos mediante productos seleccionados?
- ¿Qué se ajusta más a la situación: modelo supervisado o no supervisado?

Hipótesis: Si modificamos dinámicamente el orden de los pasillos dentro de un departamento dependiendo las selecciones de productos que el usuario vaya haciendo, entonces se reducirá un 25% el esfuerzo necesario.

Un claro ejemplo es: si quiero hacer una carne asada, y nunca he hecho una, al seleccionar la carne se me recomendará el pasillo de carbones. Estos departamentos tienen correlación entre sí por el historial de compras y patrones ocultos.

Al ordenar dinámicamente la posición de los pasillos dentro de la página de un departamento basado en los productos que tengo el carrito actualmente y la probabilidad del siguiente pasillo a visitar se reducirá el tiempo requerido para completar un pedido.

Optimizar dinámicamente la organización y agrupación de los pasillos y departamentos, basándose en los hábitos de compra y la correlación entre productos, reducirá significativamente el tiempo y el esfuerzo que los usuarios requieren para completar sus compras.

Metodología formal

La metodología propuesta tiene como objetivo validar si el reordenamiento dinámico de pasillos, basado en un modelo de aprendizaje supervisado, puede reducir el esfuerzo que realiza un usuario al llenar su carrito de compras en línea. Para ello, se definen variables clave, se prepara un entorno de simulación sobre datos históricos reales y se evalúa el impacto del modelo mediante métricas de clasificación y pruebas de eficiencia. A continuación, se describen los pasos que guiarán este proceso.

1. Definición de variables y diseño del estudio.

Se definirán tanto la variable dependiente como las de tratamiento y las covariables necesarias para evaluar el modelo y su impacto sobre la experiencia del usuario. Estas variables se especificarán en una sección posterior para mantener una estructura clara y coherente del documento.

2. Preparación y enriquecimiento del dataset.
   a. Obtener los archivos order_products_prior.csv, orders.csv, aisles.csv, y products.csv.
   b. Crear una nueva columna en la tabla aisles para asignar un número de orden a cada pasillo, permitiendo comparar recorridos.
   c. Establecer un identificador numérico para cada pasillo para calcular esfuerzo entre productos.
   d. Limpiar los datos: eliminar pedidos incompletos, eliminar duplicados, asignar pasillos a productos no etiquetados, normalizar formatos.

3. Análisis exploratorio.

a. Visualizar secuencias típicas de pasillos recorridos por los usuarios.
b. Analizar la variabilidad del orden de compra según la hora del día.
c. Calcular el esfuerzo promedio actual de los usuarios (sin modelo).
d. Identificar relaciones entre pasillos frecuentemente visitados de forma consecutiva.

4. Entrenamiento del modelo de clasificación.
   a. Convertir cada orden de compra en una secuencia de pasillos visitados.
   b. Entrenar un modelo de clasificación supervisado que, dado el historial parcial de pasillos del carrito y la hora del día, prediga cuál será el siguiente pasillo más probable.
   c. Asegurar la generalización del modelo con técnicas de implementación y validación.

5. Implementación de la lógica de reordenamiento.
   a. Simular el llenado del carrito en dos condiciones:
   b. a) con pasillos en orden fijo (baseline actual),
   c. b) con el orden dinámico sugerido por el modelo, donde al agregar un producto se recomienda el siguiente pasillo más probable.
   d. Para cada simulación, calcular el esfuerzo acumulado.

6. Evaluación de resultados:
   a. Evaluar el modelo con métricas de clasificación.
   b. Comparar el esfuerzo promedio en ambos escenarios (con y sin modelo).
   c. Realizar una prueba estadística (Prueba T de muestras pareadas) para determinar si la reducción del esfuerzo es estadísticamente significativa.

Pruebas y experimentos iniciales

Con el fin de validar la viabilidad de nuestra hipótesis y sentar las bases para el desarrollo del modelo de machine learning, se realizaron pruebas preliminares utilizando un subconjunto del dataset de Instacart. Estas pruebas tienen como objetivo observar si el reordenamiento dinámico de pasillos, basado en los productos que el usuario va agregando al carrito, puede efectivamente reducir el esfuerzo de navegación durante la compra. A continuación, se detallan las variables consideradas y los resultados obtenidos en esta fase inicial.

<u>Variable dependiente.</u>

Esfuerzo del usuario: medido como la suma de las distancias absolutas entre pasillos dentro de un mismo departamento, en el orden en que se agregan productos al carrito.

<u>Variable de tratamiento.</u>

Reordenamiento dinámico de pasillos:

- Valor 0: el orden de los pasillos es estático (baseline).
- Valor 1: el sistema reorganiza los pasillos en tiempo real según los productos en el carrito.

<u>Covariables.</u>

- Hora del día: mañana, tarde o noche (puede influir en los hábitos de compra).
- Número total de productos en la compra: más productos podrían implicar más esfuerzo.
- Tipo de usuario: nuevo vs. recurrente.
- Categoría de productos seleccionados: algunos departamentos pueden tener layouts más complejos.
- Frecuencia de compra de los productos: productos comunes vs. raros.

Para dar inicio a la experimentación, se generará una columna adicional en el dataset resultante del merge entre todo los datasets. Esta nueva columna representará el número de orden asignado a cada

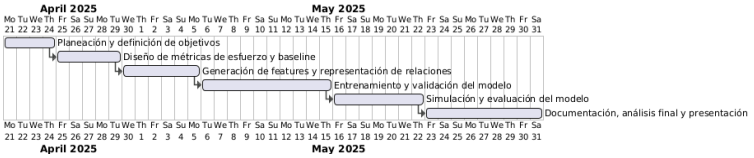
pasillo dentro de su respectivo departamento. Inicialmente, este orden será aleatorio, simulando un layout base sin optimización. A partir de esta estructura, se analizarán múltiples sesiones de compra simuladas, observando cómo los usuarios se desplazan entre pasillos y departamentos conforme agregan productos al carrito. Este enfoque permitirá medir el esfuerzo de navegación en distintos escenarios de ordenamiento y servirá como base para comparar el impacto de estrategias de reordenamiento dinámico en la experiencia de compra.

Este nuevo dataset enriquecido, que incluye la columna con el orden asignado a cada pasillo dentro de su departamento, también servirá como base para entrenar un modelo de aprendizaje supervisado de clasificación. En este contexto, cada fila del dataset puede representar una instancia en la que el sistema debe decidir la posición óptima de un pasillo, dadas ciertas condiciones como la hora del día, los productos ya agregados al carrito, el departamento al que pertenece el pasillo y la frecuencia histórica de compra. La variable objetivo será la posición ideal del pasillo (aisle_order), mientras que las variables predictoras incluirán tanto características contextuales como comportamentales. De esta manera, el modelo podrá aprender patrones que permitan anticipar qué pasillos deben mostrarse primero para reducir el esfuerzo del usuario, y así optimizar dinámicamente la experiencia de navegación en el supermercado en línea.

Cronograma de trabajo para las siguientes fases.

El diagrama de Gantt mostrado a continuación muestra la planificación del proyecto.

En general se identifican seis etapas principales: planeación y definición de objetivos, diseño de métricas para medir el impacto de la solución, generación de features y representación de relaciones, entrenamiento y validación del modelo de aprendizaje, simulación y evaluación del layout propuesto, y finalmente, la documentación, análisis final y presentación de resultados.



Referencias.

1. Arboleda, F.J.M., Garani, G., Correa, A.F.A. Supermarket Product Placement Strategies Based on Association Rules (2024) IAENG International Journal of Computer Science, 51 (6), pp. 650-662.

2. Saberi, Z., Hussain, O. K., & Saberi, M. (2023). Data-driven personalized assortment optimization by considering customers' value and their risk of churning: Case of online grocery shopping. *Computers & Industrial Engineering, 184*, 109328. https://doi.org/10.1016/j.cie.2023.109328

3. Liashenko, O., & Yakymchuk, B. (2022). Modeling the Optimal Grocery Store Trading Area Using Machine Learning Methods. *Proceedings of the Information Technology and Implementation (IT&I-2022)*, CEUR Workshop Proceedings, 3347, 325–331. https://ceur-ws.org/Vol-3347/Short_3.pdf

4. Boyer, K. K., & Hult, G. T. M. (2005). Customer behavior in an online ordering application: A decision scoring model. *Decision Sciences, 36*(4), 569–602. https://doi.org/10.1111/j.1540-5414.2005.00103.x

5. Pillai, J., Vyas, O.P., Muyeba, M. (2013). HURI – A Novel Algorithm for Mining High Utility Rare Itemsets. In: Meghanathan, N., Nagamalai, D., Chaki, N. (eds) Advances in Computing and Information Technology. Advances in Intelligent Systems and Computing, vol 177. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-31552-7_54

6. Tarigan, U., Tarigan, U. P. P., Rahman, I. H., & Rizkya, I. (2018). Design of facility layout with lean service and market basket analysis method to simplification of service process in the supermarket.

MATEC Web of Conferences, 197, 14006. https://doi.org/10.1051/matecconf/201819714006

7. Joe, T., Sreejith, R., & Sekar, K. (2019). Optimization of store layout using market basket analysis. International Journal of Recent Technology and Engineering (IJRTE), 8(2), 6459–6463. https://doi.org/10.35940/ijrte.B2207.078219

8. García-Magariño, I., Lloret, J., & Pawar, P. (2021). An E-Business Event Stream Mechanism for Improving User Tracing Processes. Computers, Materials & Continua, 69(1), 1143–1158. https://doi.org/10.32604/cmc.2021.014278

9. Budaraju, Raja Rao & Jammalamadaka, Sastry. (2025). Crowd Distance Induced Multi Objective Binary Salp Swarm Optimization Algorithm for Mining High Frequency and Utility Itemsets. SN Computer Science. 6. 1-14. 10.1007/s42979-025-03725-8.

10. Gan, W., Lin, J. C., Fournier‐Viger, P., Chao, H., Hong, T., & Fujita, H. (2018). A survey of incremental high‐utility itemset mining. WIREs Data Mining and Knowledge Discovery, 8(2). https://doi.org/10.1002/widm.1242