Paralelización del algoritmo de Random Forest para la optimización de tiempo de entrenamiento en modelo de cross/upselling para capañas de marketing. ⋆

Omar Vigueras García[1]

Universidad Politécnica Metropolitana de Hidalgo, México 213220126@upmh.edu.mx
https://www.upmetropolitana.edu.mx

Abstract. El proyecto de esta materia forma parte del proyecto: "Propuesta de algoritmo de IA para la optimización del ROI de las campaña de marketing de conversión y awareness del periodo Enero-Agosto 2023 caso KIO", el cual se complementa con el objetivo de optimizar el modelo de entrenamiento del algoritmo de Random Forest para obtener las características principales de los clientes que tienden a comprar un determinado producto. Como resultado del proyecto se pretende segmentar el paquete de servicios para obtener un mayor enfoque de campañas de marketing hacia los clientes con base en la probabilidad de compra.
Este proyecto evalúa las partes del algoritmo con el objetivo de presentar las partes paralelizables que permitan tener una reducción de clasificación de datos y tiempo de entrenamiento para lograr una optimización de recursos computacionales. Se entrena utilizando la base de datos de oportunidades de venta ganadas y generadas en la compañía durante el periodo de 2019-2022.
El objetivo de la aplicación de los algoritmos de clasificación es identificar las características de clientes que tienden a adquirir un producto o servicio dentro del portafolio para generar una estrategia de venta cruzada dentro de los clientes existentes e identificar mediante características los prospectos potenciales con una alta probabilidad de compra para ofrecer nuevos servicios.

Keywords: Análisis predictivo · modelos de clasificación · ciencia de datos en marketing.

1 Esquema general del proyecto

1.1 Técnicas de modelado de análisis predictivo en marketing con ciencia de datos.

Como componente principal en el desarrollo de una campaña de marketing, el entendimiento del mercado es vital. Ya que las estrategias a desarrollar deben

⋆ Caso KIO

de enfocarse a públicos, sectores, estilos de vida y necesidades específicas. Aún dentro de un gran portafolio de productos dentro de una organización específica, se llegan a crear segmentos diferentes y específicos a los cuales hay que prestar atención para poder tener el máximo rendimiento posible en las campañas.

Cuando se habla de un portafolio de más de 2,000 productos servicios diferentes, los segmentos no son tan fáciles de identificar, para lo cual se deben implementar modelos que nos ayuden a identificarlos por nosotros y los cuales se actualicen constantemente para identificar los nuevos comportamientos del mercado.

Miller en su libro de Marketing Data Science (2015)[1] Hace referencia a un nuevo tipo de marketing ya que en un mundo transformado por la tecnología de la información los datos gobiernan el día y que la ciencia de datos es considerada la nueva estadística. A través del uso de información disponible de la base de datos de la empresa se pueden generar modelos de ciencia de datos que permitan la comprensión del mercado, identificar características clave de los clientes y generar análisis predictivos de compra de productos.

Los modelos de regresiones lineales, logística o de clasificaciones como k-means, árboles de decisión son buenas alternativas para segmentar el mercado actual y estimar la probabilidad de una respuesta favorable de compra ante nuevos segmentos similares que no estén dentro de nuestra cartera usando la información que la propia compañía genera.

Otras contribuciones dentro del campo del marketing pueden nutrir la aplicación de la IA en el entendimiento del mercado, tales como Anand y Büchner(2002)[2] en el libro Database marketing and web mining, donde muestran una identificación de prospectos para la venta cruzada de una lista de clientes existentes, los métodos bayesianos en marketing, revisados por Rossi, Allenby, McCulloch en su libro Bayesian Statistics and Market-ing(2004)[3]

1.2 Selección y arquitectura de datos para la predicción de compra

Para modelar un algoritmo que prediga las caraterísticas de los clientes que tienden adquirir un determinado producto o servicio, debe tomarse en cuenta los valores del cliente que pueden obtenerse tanto de forma interna como externa, tales como el número de empleados, tamaño de la empresa, nivel de facturación, estado, etc. De esta forma el modelo podrá calcular la probabilidad de compra de nuevos clientes al comparar el patrón de venta que las empresas por sus características han tenido en la compañía. La Tabla 1 brinda un ejemplo de los tipos de datos que se toma en cuenta para la construcción de la base de datos del modelo.

Arquitectura de datos Para la construcción del modelo de datos se plantea un proceso de 3 fases: (verFig. 1).

1 Extracción de información Extracción de los datos del CRM (Customer relationship management) de acuerdo a las características e historial de venta de los productos y servicios de la compañía tomando en cuenta el historial de los años 2019 al 2022.

2 Clasificación de categorías de productos Se realiza una clasificación de los productos de acuerdo a un estándar de etiquetado de la compañía, por lo que se busca reclasificar el historial de productos de los últimos 4 años.

3 Entrenamiento y despliegue del modelo Se ejecuta la fase de entrenamiento y el modelo de paralelización para la optimización del algoritmo.

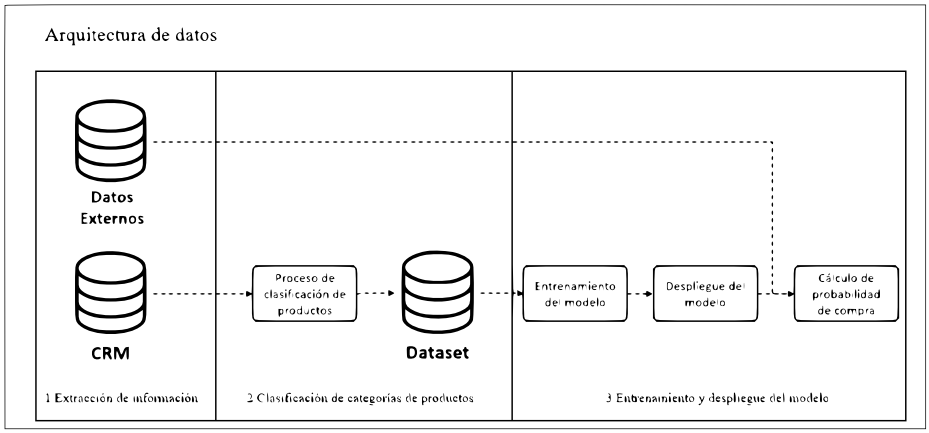

Fig. 1. Arquitectura de datos

1.3 Construcción del algoritmo de Árbol de Decisión

Selección de parametros Para la construcción del modelo se selecciona un dataset con 397,924 registros en 9 columnas las cuales representan del tipo de venta del producto. Se define la columna "Venta producto 1" como la variable dependiente del modelo el cual contiene un valor booleano, de tal manera que si el producto fue vendido se registra como 1 de lo contrario 0. Adicional para entrenar el modelo se divide el 80% del conjunto de datos para entrenar (training) el modelo y el 20% restante para prueba (testing) del modelo.

Table 1. Tipos de datos seleccionados del modelo

Nombre	Tipo de datos	Tipo de variable
Empresa	**String**	Independiente
Origen del lead	**String**	Independiente
Renovación	**Bool**	Independiente
Contacto	**Bool**	Independiente
Sector	**String**	Independiente
País	**String**	Independiente
División	**Integer**	Independiente
Industria	**Integer**	Independiente
Familia de producto	**Integer**	Independiente
Ventas Clasificación	**Float**	Independiente
Número Empleados	**Integer**	Independiente
Tamaño de empresa	**String**	Independiente
Venta	**Integer**	Dependiente

Paralelización del Modelo de clasificación

Implementación del modelo de paralelización) Con el objetivo de optimizar el algoritmo en la fase de preparación y entrenamiento, se propone un modelo paralelización de acuerdo al número de procesadores del equipo local, realizando particiones en las lista de los datos al momento de realizar la clasificación de los datos (fase 2) (verFig. 2) y posteriormente en la fase de entrenamiento una paralelización por hiperparámetros para un clasificador tipo Random Forest. Definiendo un numero de árboles de 397, un criterio: Gini y 3 hilos.

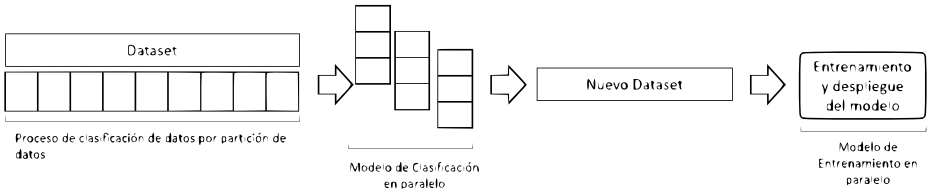

Fig. 2. Arquitectura de datos

Resultados) Al evaluar nuestro Random forest sin optimizar el equipo de computo, el programa se concluye en 253 min. sin embargo al distribuir las cargas de trabajo de los 397 árboles en 3 hilos, el tiempo se reduce a 0.007496 minutos.

Evaluación incrementando el número de árboles) Con el objetivo de medir la eficacia del modelo de paralelización se genera una prueba adicional incrementando el número de árboles del modelo a 997, teniendo como resultado un total de: 0.01004 minutos o 0.60 segundos. Adicional una evaluación con 2,404 árboles, los resultados se muestran en la Tabla 2

Table 2. Resultados de tiempo de entrenamiento

Tipo de entrenamiento	Número de árboles	Tiempo en segundos
Normal	397	253 minutos
Distribuido en 3 Hilos	397	0.007496 minutos
Distribuido en 7 Hilos	997	0.009714 minutos
Distribuido en 7 Hilos	2,404	0.009460 minutos

References

1. Miller, T. W.: Marketing Data Science. 1st edn. Pearson Education, Old Tappan New Jersey (2015)
2. Anand, S. S. and A. G. Büchner: Database marketing and web mining. In W. Klösgen and J. M. ytkow (eds.), Handbook of Data Mining and Knowledge Discovery, Chapter 46.1, pp. 843–849. Oxford: Oxford University Press.(2002)
3. Rossi, P. E., G. M. Allenby, and R. McCulloch. Bayesian Statistics and Marketing.New York: Wiley (2005).