



Excelencia que trasciende

DELVALLE
GRUPO EDUCATIVO

Proyecto 2 Resultados Iniciales Data Science

**Hugo Rivas - 22500
Alexis Mesias - 22562
Julio Lemus - 22461**

Resumen

El propósito de esta revisión es identificar y analizar los algoritmos de aprendizaje automático y las arquitecturas predominantes para predecir la recompra en entornos no contractuales (e-commerce). Se sistematizan cinco familias metodológicas: (1) modelos probabilísticos BTYD (BG/NBD y Pareto/NBD), (2) modelos de clasificación supervisada con variables RFM/IPD y atributos contextuales, (3) modelos de supervivencia para tiempo-al-evento (tiempo a la siguiente compra), (4) modelos secuenciales profundos, y (5) uplift modeling cuando el objetivo final es la optimización de campañas y no solo la predicción de comportamiento. La evidencia en revistas indexadas muestra que los GBMs y ensambles superan con frecuencia a líneas base lineales; los BTYD son estándares sólidos de referencia cuando el histórico transaccional está bien fechado; los enfoques de supervivencia mejoran cuando el timing es crítico; y los modelos de uplift son superiores para decidir a quién contactar. Se discuten implicaciones prácticas para el flujo del equipo (RFM, IPD, split temporal, métricas y lift).

1. Introducción

En un comercio electrónico típico, una porción significativa de clientes compra una sola vez; por ello, anticipar quién repetirá compra y cuándo lo hará permite activar retención y asignar presupuesto con mayor retorno. El equipo plantea cuestionar líneas base como “predecir siempre la clase mayoritaria” o reglas RFM simples, utilizando validación temporal y métricas adecuadas. Esta revisión ubica el problema en la literatura de CRM y analítica de marketing y organiza la evidencia por familias de modelos.

2. Metodología de la revisión

Se priorizaron artículos de revistas indexadas en marketing analítico, OR/analytics y HCI/ML aplicados, que reportan predicción de recompra o tareas directamente afines (compra futura, retención, time-to-repurchase, respuesta a marketing y targeting). Se incluyeron trabajos seminales (BTYD), comparativos con GBMs/ensambles, propuestas con supervivencia y profundo secuencial, y revisiones/benchmarking de uplift. Se favoreció investigación con datos reales y evaluación mediante AUC, F1, métricas de supervivencia y/o lift.

3. Marco conceptual y familias de algoritmos

3.1. Modelos BTYD: BG/NBD y Pareto/NBD

Los modelos Buy-'Til-You-Die (BTYD) cuantifican probabilidad de “seguir activo” y número esperado de compras futuras a partir de recencia y frecuencia. BG/NBD (Fader, Hardie y Lee) se consolidó como alternativa computacionalmente más simple al Pareto/NBD (Schmittlein, Morrison y Colombo), manteniendo precisión competitiva en entornos no contractuales. Estos modelos son referencia habitual para pronóstico de recompra y CLV y, por tanto, constituyen un baseline probabilístico valioso frente a reglas RFM. (brucehardie.com)

Asimismo, evidencia reciente en marketing muestra que incorporar regularidad temporal mejora la capacidad predictiva respecto a enfoques basados solo en RFM, sin perder interpretabilidad. ([ScienceDirect](#))

Implicación: En el proyecto, BG/NBD puede operar como baseline adicional al conjunto de reglas RFM y la línea base de mayoría, reforzando la comparación empírica propuesta por el equipo.

3.2. Clasificación supervisada con RFM/IPD y variables contextuales

Los modelos discriminativos con ingeniería de variables (R, F, M, IPD/ritmo, temporada, canal, cupones, devoluciones, país/cohorte) suelen liderar en precisión cuando se capturan no linealidades e interacciones. La evidencia en revistas y estudios aplicados sugiere que ensambles y GBMs (Random Forest, XGBoost/LightGBM/CatBoost) superan a la regresión logística y a árboles simples en AUC/F1 para repetición de compra y conversión. ([IDEAS/RePEc](#))

Adicionalmente, el uso de señales de interacción de bajo nivel aporta ganancia incremental sobre logs puramente transaccionales en la predicción de repeat purchase, según evidencia reciente en International Journal of Human–Computer Interaction. ([Taylor & Francis Online](#))

Implicación: El flujo del equipo ya contempla RFM e IPD; incorporar variabilidad del IPD, estacionalidad y efectos de promociones/cupones es consistente con la literatura; evaluar LightGBM/XGBoost/CatBoost contra la logística baseline es adecuado para cuestionar la línea base.

3.3. Modelos de supervivencia (tiempo a la siguiente compra)

Cuando importa cuándo volverá a comprar el cliente, los modelos de supervivencia tratan naturalmente la censura. La literatura en PLOS ONE muestra que técnicas de supervivencia (p. ej., Cox, Random Survival Forests y variantes) permiten aprender distribuciones individuales de comportamiento de compra y son útiles para políticas de timing de marketing. ([PLOS](#))

Estudios aplicados recientes reportan mejoras al combinar RSF con segmentación/ensembles sobre grandes CRM para repurchase time. ([MDPI](#))

Implicación: Además de predecir “repetirá / no repetirá”, modelar el tiempo (hazard/riesgo) permite disparar comunicaciones unos días antes del IPD esperado por segmento, alineado con el uso de IPD que ya se reportó en el EDA del equipo.

3.4. Modelos secuenciales profundos

RNN/LSTM/GRU han probado ser competitivos en analítica de marketing y respuesta, al explotar el orden de eventos y reducir la dependencia de feature engineering. En Journal of Interactive Marketing, un estudio comparativo muestra que un LSTM supera a una gran mayoría de modelos tradicionales basados en features. Este enfoque es particularmente útil con secuencias densas (clics, vistas, sesiones) y datos de panel. ([IDEAS/RePEc](#))

Implicación: Si el dataset del proyecto incluye registros de interacción o secuencias temporales ricas, un LSTM/GRU con atención podría evaluarse como modelo complementario; de lo contrario, el costo/beneficio puede ser menor que GBMs bien diseñados.

3.5. Uplift modeling para targeting de retención

Si la predicción se usará para decidir a quién contactar (cupón, email), el objetivo real no es la probabilidad de recompra sino el efecto causal individual (incremento por tratamiento). El Journal of Interactive Marketing presenta un benchmark sistemático donde métodos de uplift y HTE (treatment effects) se comparan en datos reales y sintéticos, con recomendaciones sobre evaluación (curvas Qini/uplift y políticas de targeting). ([SAGE Journals](#))

Implicación: Si el curso permite extender objetivos hacia retención accionable, incorporar un experimento A/B o un histórico de campañas haría procedente evaluar árboles/bosques de uplift o meta-aprendices (T-/X-/R-learner) y medir uplift por deciles.

4. Discusión: alineación con el caso del curso

El grupo ya documentó: (i) variables RFM e IPD, (ii) líneas base (mayoría, reglas RFM), (iii) split temporal, y (iv) métricas ROC-AUC/PR-AUC, F1 y lift por deciles. La literatura respalda:

- BTYD (BG/NBD y Pareto/NBD) como baseline probabilístico adicional y explicativo. ([brucehardie.com](#))
- GBMs/ensembles con *features* derivadas de RFM + IPD + temporada + promociones como candidatos principales para superar líneas base. ([IDEAS/RePEc](#))
- Supervivencia cuando interesa la activación por timing (hazard de recompra). ([PLOS](#))
- Profundo secuencial si existe traza densa de eventos (beneficio marginal sobre GBMs). ([IDEAS/RePEc](#))

5. Selección de algoritmo a utilizar

Se seleccionará LightGBM como modelo principal para la predicción de recompra, dado su desempeño consistente en datos tabulares con variables RFM e IPD, su capacidad para capturar no linealidades e interacciones sin requerir ingeniería de características extensa y su eficiencia computacional; como líneas de contraste y respaldo, se incluirán Regresión Logística, por su interpretabilidad, rapidez y utilidad como piso metodológico y de calibración, y CatBoost como alternativa robusta cuando predominan variables categóricas, de modo que se compare rendimiento y estabilidad entre enfoques lineales y de gradient boosting bajo el mismo esquema de validación temporal y las mismas métricas (ROC-AUC, PR-AUC, F1 y lift por deciles).

6. Conclusiones

La evidencia en revistas indexadas converge en que: (a) BG/NBD y su familia siguen siendo estándares para pronóstico de actividad y conteo en entornos no contractuales; (b) los GBMs con feature engineering pertinente son fuertes candidatos para clasificación de recompra y suelen superar

líneas base lineales; (c) los modelos de supervivencia agregan valor cuando el tiempo de recompra es relevante para la orquestación de campañas; (d) los modelos secuenciales profundos ofrecen ventajas en presencia de secuencias de interacción densas; y (e) si la meta es optimizar el targeting, la métrica y el modelo correctos pertenecen al marco de uplift. Se recomienda que el equipo compare Logística (baseline interpretable), Random Forest/GBMs (modelos principales), BG/NBD (baseline probabilístico) y, si el alcance lo permite, un modelo de supervivencia para timing.

Referencias

- Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). “Counting your customers” the easy way: An alternative to the Pareto/NBD model. *Marketing Science*, 24(2), 275–284. <https://doi.org/10.1287/mksc.1040.0098> (brucehardie.com)
- Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting your customers: Who are they and what will they do next? *Management Science*, 33(1), 1–24. <https://doi.org/10.1287/mnsc.33.1.1> (Semantic Scholar)
- Reutterer, T., Platzer, M., & Schröder, N. (2021). Leveraging purchase regularity for predicting customer behavior the easy way. *International Journal of Research in Marketing*, 38(1), 194–215. <https://doi.org/10.1016/j.ijresmar.2020.09.002> (ScienceDirect)
- Jin, P., Wang, H., Wang, P., & Yang, J. (2021). Using survival prediction techniques to learn consumer repurchase behavior. *PLOS ONE*, 16(5), e0251623. <https://doi.org/10.1371/journal.pone.0251623> (PLOS)
- Larivière, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, 29(2), 472–484. <https://doi.org/10.1016/j.eswa.2005.04.043> (ScienceDirect)
- Martínez, A., Schmuck, C., Pereverzyev Jr., S., Pirker, C., & Haltmeier, M. (2020). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*, 281(3), 588–596. <https://doi.org/10.1016/j.ejor.2018.04.034> (IDEAS/RePEc)
- Kuric, E., Puskas, A., Demcak, P., & Mensatorisova, D. (2024). Effect of low-level interaction data in repeat purchase prediction task. *International Journal of Human–Computer Interaction*, 40(10), 2515–2533. <https://doi.org/10.1080/10447318.2023.2175973> (Taylor & Francis Online)
- Sarkar, M., & De Bruyn, A. (2021). LSTM response models for direct marketing analytics: Replacing feature engineering with deep learning. *Journal of Interactive Marketing*, 53, 80–95. <https://doi.org/10.1016/j.intmar.2020.07.002> (IDEAS/RePEc)
- Rößler, J., & Schoder, D. (2022). Bridging the gap: A systematic benchmarking of uplift modeling and heterogeneous treatment effects methods. *Journal of Interactive Marketing*, 58, 1–22. <https://doi.org/10.1177/10949968221111083> (SAGE Journals)