IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication 20 de Junio de 2023, date of current version 20 de Junio de 2023.

Digital Object Identifier 10.1109/ACCESS.2022.Doi Number

Modelo predictivo de transacciones Evertec Placetopay S.A.S

L.C. Torres López[1], E.M. Jiménez Marín[2]

[1] Departamento de Ingenieria, Universidad de Antioquia, Medellín, CO

[2] Departamento de Ingenieria, Universidad de Antioquia, Medellín CO

Corresponding author: L.C Torres López (laura.torresl@udea.edu.co). E.M Jiménez Marín (esteban.jimenezm@udea.edu.co).

ABSTRACT Evertec Placetopay S.A.S es una empresa de tecnología que focaliza sus servicios en una plataforma transaccional de comercio electrónico y pagos no presenciales, en Colombia y, en otros países de Latinoamérica. Por la naturaleza de sus servicios, la empresa posee un alto volumen de datos acerca del comportamiento transaccional de sus usuarios y comercios; sin embargo, en la actualidad no cuenta con modelos de machine learning robustos y refinados. Dicha necesidad se atiende en el marco de este proyecto que tiene por objetivo crear un modelo predictivo de transacciones, partiendo de un esquema no supervisado, con la segmentación de clientes (análisis cluster), para, finalmente, crear un modelo de regresión que prediga cuánto transará un comercio. En este documento, se consignan los principales hallazgos de una exploración inicial de los datos, se describe el proceso implementado, se definen posibles métricas o medidas de desempeño, y, se informan otros referentes que han trabajado el mismo tema.

INDEX TERMS Electronic Commerce, Machine Learning, Pattern Clustering, Predictive Models,

I. INTRODUCTION

Es indudable que la pandemia por COVID-19 marcó un antes y un después en el mundo. Diferentes sectores, como el educativo, el laboral, el económico, entre otros, tuvieron un efecto y un comportamiento inusial en este este período. Como es de esperarse, el comercio electrónico también vivió un cambio a raíz del confinanmiento, específicamente, en Colombia, las transacciones electrónicas se posicionaron mostrando un crecimiento constante durante el 2020 y 2021. Según la Cámara Colombiana de Comercio Electrónico [2] el aumento de transacciones en líne en el año 2020, con respecto al año 2019, fue de 79,4%, y, el aumento de ventas digitales en el tercer trimestre del año 2021, con respecto al mistro trimestre de 2019, fue de 79,6%.

Estas cifras confirman algo que es evidente: la cantidad de información sobre las ventas y transacciones digitales también incrementó. De esta manera, empresas como Evertec Placetopay S.A.S, que ofrecen soluciones tecnológicas que habilitan canales de venta y cobro en línea para comercios grandes y pequeños del país, adquirieron un alto volumen de información acerca del comportamiento transaccional de sus usuarios y comercios. Lamentablemente, no todas las empresas del comercio digital, estaban preparadas para tratar dichos datos.

En el caso de Evertec Placetopay, no existe una madurez en la infraestructura ni un proceso establecido para proyectos de tratamiento de datos y machine learning. Por consecuente, la empresa no cuenta con modelos predictivos de regresión o clasificación robustos y refinados, de igual manera, sus comercios no se encuentran clasificados ni segmentados. Esta situación tiene una incidencia directa en el desarrollo de la compañía; debido a que, el core del negocio está en sus servicios y soluciones transaccionales, y, evidentemente, son los comercios los que consumen estos servicios. Por ende, la atracción y cierre de contratos con antiguos y nuevos clientes es fundamental para la subsistencia y crecimiento de la organización. En el ámbito de la negociación se vuelve fundamental contar con modelos de análisis de datos que orienten la toma de decisiones particulares, especialmente, modelo de predicción o pronóstico,

En línea con lo anterior, en el marco de este proyecto se propone crear un modelo regresion para predecir cuánto transará un comercio. De forma complementaria, se plantea un análisis cluster para segmentar los comercios en grupos claramente diferenciados que nos informen de un comportamiento particular. Con esto, esperamos proveer una base sólida para el refinamiento de modelos predictivos en la organización.

II. REFERENTES

Antes de proceder con el tratamiento de datos y construcción de los modelos, se exploraron las investigaciones y estudios realizados por otros autores, con la intención de hallar un punto de partida o referencia frente a las técnicas, metodologías y resultados frecuentes.

A, Kumar [1], creó un modelo cluster para la segmnentación de clientes de un mall comercial, y, de esta manera lograr que se ofrezcan servicios y promociones diferenciadas a los consumidores. Su metodología se basó en cinco pasos: 1. Entendimiento de la necesidad del negocio. 2. Exploración y comprensión de los datos. 3. Prueba y combinación de diferentes hiperparametros. 4. Selección del mejor modelo incluyendo su validación. 5. Implementación del modelo. Siguiendo estos pasos, Kumar, se centró en el algoritmo K-means; aunque este no es el único algoritmo utilizado para este tipo de análisis (también está el K-medoids y Fuzzy C-means), si es de los más utilizados. Igualmente, como es frecuente, dicho autor se basa en el conocido gráfico de codos para definir el número óptimo de grupos. Como resultado, el autor creó dos modelos de segmentación: el primero, en función de la edad y el puntaje de gasto de los clientes; el segundo en función de los ingresos anuales y puntaje de gasto de los mismos. Para el primer modelo se definió un número óptimo de 4 clusters donde cada uno de ellos indicaba si la persona es de avanzada o corta edad y si tenía una mayor o menor tendencia a gastar dinero. Y, Para el segundo, se definió una cantidad de 5 clusters los cuales daban un indicio de los ingresos del individuo teniendo en cuenta su edad. Ambos agrupamientos posibilitaron la toma de decisiones estratégicas, tales como campañas de fidelización o formulación de planes de ventas dirigidos a un segmento de la población en específico.

En su investigación, K, Tabianan, S, Velu & V, Ravi [3], se acercan más a nuestro tema de trabajo: el comercio electrónico. En su caso, se centraron en determinar el comportamiento de compra de los clientes a través de un análisis cluster. Igual que el estudio anterior, se basaron en el algoritmo K-means. Para validar el desempeño del algoritmo de agrupamiento, los autores contrastaron la segmentación obtenida a partir del modelo cluster con el histórico de compras de los clientes. Como resultado, al comparar el histórico de compra de los clientes para cada uno de los cluster resultantes, se identificó que efectivamente cada segmento cuenta con un patrón de consumo definido, ya sea la compra o solo la vista del producto como tal, captando así, la desigualdad que se presenta entre los segmentos de baja y alta rentabilidad en la categoría de producto.

L, Cruz & R, Arano [4], se plantearon como objetivo en su investigación, obtener estimaciones y pronósticos de valores futuros, del comercio electrónico en México, a partir de la información histórica, implementando dos modelos no lineales, el primero exponencial y el segundo logístico. Como resultado, los autores observaron que existe una tasa de crecimiento palpable tanto para el número de usuarios como la utilización del comercio electrónico en las ventas, notando así un aumento importante año tras año de manera geométrica en el mediano plazo que supera a las expectativas de un crecimiento lineal, y en a largo plazo generará un crecimiento paulatinamente decelerado, obedeciendo otro modelo, tal como el logístico.

Por su parte, L, Martínez, M, Otálora, L, Prada & L, Skinner [5], se plantearon el objetivo de predecir el incremento de ventas de un comercio local, basados en el histórico de promoción y publicidad de 5 productos de la empresa Aqua. Inicialmente, utilizaron un modelo de regresion logística binaria, para reducir el número de variables de entrada, y, posteriormente implementar un modelo de regresion de Poisson. Las autoras dividieron su base de datos en datos de entrenamiento (80%) y datos de prueba (20%). Para la evaluación del modelo de Poisson, con los datos de prueba, se procedió a identificar el RMSE (Error de Raíz Cuadrada Media).

Los estudios referenciados apuntan a modelos diferentes, en los dos primeros se observa un enfoque al análisis cluster, y en los dos últimos a modelos de regresion lineal o no lineal. Esto se debe, a que en el presente trabajo se propone implementar ambos modelos, considerando que, a partir de un esquema no supervisado, se llegue a una predicción bajo un esquema supervisado. Por tanto, se exploran ambas aristas.

III. EXPERIMENTOS

A. METODOLOGÍA

Para el desarrollo de esta monografía se consideran tres etapas comprendidas por el pretratamiento de los datos suministrados por Evertec Placetopay S.A.S, clustering o clasificación de los comercios y el planteamiento de un modelo de regresión con el fin de predecir la variable de interés.

Para el pretratamiento del dataset se hizo una exploración preliminar en la cuál se procedió con una revisión de la calidad de los datos, la eliminación y creación de variables que se consideraron pertinentes, imputación de valores nulos por medio del algoritmo de vecinos más cercanos (KNN) y valores más frecuentes (moda) para las variables numéricas y categóricas respectivamente, seguido de una escalización MinMax y una eliminación de datos atípicos a través del método LOF.

Durante el proceso de agrupamiento, se hizo uso del diagrama de codo para tener una referencia de la cantidad de clusters óptimos (k) en función de los parámetros definidos cómo el error y máximo de iteraciones. Para el clusteríng se usó el algoritmo correspondiente a K-means.

Finalmente, luego de obtener un vector de etiquetas, se uso esta clasificación de los comercios para realizar una regresión lineal con el fin de predecir el valor de la característica objetivo 'Monto mes USD'.

La base de datos fue suministrada por la empresa Evertec Placetopay, cuenta con 108512 instancias y un total de 26 variables o características.

B. VALIDACIÓN

Para validar el agrupamiento realizado, se definen las métricas de validación interna para clusterining como los puntajes de silueta, calinski-harabasz y davies-bouldin. Se dará prioridad al puntaje de silueta.

Para el modelo predictivo de regresión, se tendran métricas basadas en el error resultante de la diferencia del valor real y el valor predicho por el modelo como el MSE y el RMSE, acompañado además del R2. No obstante, se elige el RMSE como métrica de desempeño principal debido a su fácil interpretación para el negocio.

IV. RESULTADOS Y DISCUSIÓN

Para el agrupamiento realizado con KMeans, se obtuvo una métrica de Silueta correspondiente a 0.224741, lo cual, por su cercanía a cero, indica que hay cierta superposición o proximidad entre los grupos resultantes. Adicionalmente, las métricas de calinski-harabasz y davies-bouldin (16682.741996, y 1.694246 respectivamente) podrían indicar una baja dispersión entre los grupos y una separación regular entre los mismos.

Por lo anterior, se plantea a futuro la posibilidad de realizar modificaciones con respecto a las variables a eliminar en función de sus correlaciones y significancia estadística, además de una posible reducción de componentes que permita disminuir la dimencionalidad y pueda favorecer la separabilidad y compactación del agrupamiento.

Finalmente para el modelo de regresión lineal desarrollado, se obtuvieron los siguientes resultados para los datos de entrenamiento, test y prueba (los datos de prueba se obtuvieron al separar 10% de las instancias del dataset).

Entrenamiento:
El error MSE de test es: 1132564168253.4832
El error RMSE de test es: 1064219.9811380554
El error R2 de test es: 0.2587114465973497

Test:
El error MSE de test es: 917460338324.3258
El error RMSE de test es: 957841.499583478
El error R2 de test es: 0.24444335882846369

Prueba:
El error MSE de test es: 601031686953.1589
El error RMSE de test es: 775262.3342799254
El error R2 de test es: 0.11883478572455408

En los tres escenarios de validación tanto el RMSE, como el R2 indican un bajo rendimiento y precisión del modelo. Por su parte, una magnitud del RMSE del orden de 1e5 en adelante representa un error bastante significativo considerando que dicho valor corresponde a montos monetarios en USD. Por otro lado, un R2 inferior a 0.3 supone un ajuste pobre al modelo lineal por lo cual se debería reconsiderar otros enfoques que permitan el entrenamiento de un modelo con mejor desempeño.

Para dicho cometido, se plantea la posibilidad de entrenar modelos de regresión no lineales, máquinas de soporte vectorial e incluso árboles de decisión, acompañados de técnicas de búsqueda de hiperparámetros como gridsearch y validación cruzada para hacer uso de la información disponible en su totalidad. De esta manera se espera desarrollar un modelo que Evertec Placetopay pueda desplegar en producción y sirva como herramienta de apoyo para la toma de decisiones estratégicas.

REFERENCES

[1] A. Kumar. "Customer Segmentation of Shopping Mall Users Using K-Means Clustering". *IGI Global. 2022.*

[2] E. M. Ramírez. ¿Qué pasó con el comercio electrónico en 2021? 2021. [En línea]. Disponible en: https://www.ccce.org.co/noticias/que-paso-con-el-comercio-electronico-en-2021/

[3] K. Tabianan, S. Velu, y V. Ravi. “K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data”. *Sustainability*, ser14, n7243, pp 1-15, *2022.*

[4] L.C. Kuri y R. M. Arano. “Predicciones relativas al comercio electrónico en México mediante dos modelos de regresion no lineal”. *Asosiación Mexicana de internet.* 2017. [En línea]. Disponible en: https://www.uv.mx/iiesca/files/2017/10/04CA201701.pdf

[5] L. Y. Martínez, M. P. Otálora, L. Prada, y L. V. Skinner, Diseño de un sistema de apoyo para la toma de decisiones de marketing que modele los efectos de las promociones y la publicidad en las ventas del emprendimiento Aqua, [Tesis de pregrado], Pontificia Universidad Javeriana, Bogotá, 2022.