Predicción de aceptación de cupones para viajeros con modelos de clasificación

Davis Gereda
Escuela de Posgrado
Pontificia Universidad Católica del Perú
Lima, Perú
davis.gereda@pucp.edu.pe

Josue Mauricio
Escuela de Posgrado
Pontificia Universidad Católica del Perú
Lima, Perú
a20224110@pucp.edu.pe

Iván León
Escuela de Posgrado
Pontificia Universidad Católica del Perú
Lima, Perú
ivan.leong@pucp.edu.pe

***Abstract*—El presente trabajo realizará un estudio del uso de distintos modelos de clasificación, entre ellos Árbol de Decisión, KNN, Random Forest, XGBoost, Clasificación Logística y SVM. El principal objetivo es mostrar las ventajas que brinda el uso de estos modelos de clasificación en la aceptación de cupones por sistemas de recomendación, específicamente en usuarios de un vehículo. Con la retroalimentación de los resultados de la predicción hecha con estos modelos, se busca que los expertos en ventas puedan utilizar esta herramienta para ofrecer mejores ofertas y mejorar la publicidad de los productos recomendados para cada cliente.**

***Keywords*—Machine Learning, Modelos de clasificación, Sistemas de Recomendación**

I. Introducción

Los sistemas de recomendación se han convertido en herramientas importantes dentro de la industria del marketing, ya que con el tiempo distintas empresas de cualquier categoría hacen uso del beneficio de estos para mejorar las ventas y explorar en lo que desea o necesita un cliente. Dentro de los sistemas de recomendación surgen ciertas heurísticas con respecto al usuario que decidirá acerca de un producto recomendado. Un concepto importante es que la empatía es vital en sistemas orientados a seres humanos para la resolución de problemas, entendimiento mutuo y relaciones sostenibles [1]. Por ello los sistemas buscarán acertar las necesidades de un cliente generando una relación entre las recomendaciones realizadas por el sistema y la aceptación del cliente. Actualmente, compañías como Google, Amazon y Netflix hacen uso de sistemas de recomendación por medio de publicidad, networking y películas.

El objetivo de este estudio es mostrar las ventajas del uso de modelos de clasificación que ayuden a los sistemas de recomendación a mejorar las estrategias de recomendación y su eficiencia con respecto a la aceptación de los usuarios. Específicamente, se usará el conjunto de datos *in-vehicle coupon recommendation* [2] proporcionado por el repositorio de Aprendizaje Automático de la Universidad de California Irvine (UCI) el cual contiene 12,684 registros etiquetados de forma binaria por la variable objetivo 'Y' con valores 0 y 1. Se propone el uso de los siguientes modelos de clasificación: Árboles de decisión, Random Forest, XGBoost, Clasificación Logística, KNN y SVM. Se mostrará en tablas comparativas los resultados de la precisión y métricas de los modelos con los datos anteriormente presentados.

La estructura de este artículo es la siguiente: En la sección 2 se explicará el estado del arte, que revisará los distintos estudios realizados en la literatura. La sección 3 se encargará de mostrar el diseño del experimento, describiendo uno a uno los datos utilizados, también se presentará la metodología utilizada para el pre-procesamiento de los datos, entrenamiento y ajuste de los modelos de clasificación propuestos. En la sección 4, se muestra la experimentación y resultados, donde se presentará en tablas comparativas los resultados de nuestra experimentación, adicionalmente se expondrá la visualización de las métricas de clasificación mediante gráficos. La sección 5, encargado de la discusión de los resultados expuestos en la sección 4. Por último, se presentarán las conclusiones de nuestro artículo.

II. Estado del Arte

En la literatura, se han realizado estudios que evalúan modelos de clasificación para mejorar la predicción respecto a sistemas de recomendación. Boteju, P. [3], presenta un estudio basado en un sistema de recomendación de vehículos, que ayuda a un usuario a decidir sobre la compra de un vehículo. Concluye que los modelos de redes neuronales son más eficientes que modelos Random Forest y Regresión logística de clasificación múltiple, logrando cerca de un 96.23% de precisión en los datos de validación.

Con respecto a los datos, Tong Wang [2], introduce el modelo conjunto de reglas basado en reglas bayesianas. Logrando una precisión del 77%, mucho mejor en comparación de otros modelos de regresión como CART, Lasso y RIPPER. Tran Duc [4], hace un estudio de rendimiento de los modelos Árbol de Decisión, Random Forest, Máquina de Vectores de Soporte (SVM), Redes Neuronales de Avance (MLP), Regresión Logística, Bagging, AdaBoost y XGBoost. Para el entrenamiento de los modelos se crearon 4 subconjunto de datos de los cuales fueron sometidos a un pre-procesamiento usando 2 métodos para completar los datos faltantes, los métodos empleados fueron Mode Imputing y Random Forest Imputing. Asimismo, para el pre-procesamiento se empleó

el escalado de datos usando Min-Max solo para dos subconjuntos dejando los dos restantes sin escalado de datos. Concluye una exactitud que se encuentra en el rango de 68% a 76%. El modelo sugerido para obtener los mejores resultados de rendimiento es el modelo de ensamble Bagging con el estimador base Random Forest, esta sugerencia del estudio analizado nos brinda un importante punto de partida para iniciar nuestra experimentación en busca del mejor resultado. Enes Çelik [5], realiza una comparación de distintos modelos de clasificación con respecto a los datos de [2], teniendo el mejor rendimiento con el modelo LightGBM, un modelo basado en Gradient Boosting. Con este modelo se alcanzó un 75% de precisión, en general se concluye que los modelos de Gradient Boosting logran tiempos cortos de procesamiento, mientras que otros modelos como SVM son mucho más lentos. D. R. Hermawan [6], realiza un estudio comparativo entre modelos de clasificación de árboles de decisión para determinar la aceptación de cupones utilizando los algoritmos J48, Árboles de decisión y Random Forest. Se concluye de los estudios mencionados anteriormente que el modelo Random Forest obtuvo una precisión de 77.09%, el cual fue el modelo base propuesto por el autor. P. Pokhrel [7], al igual que [5], concluye que los modelos basados en ensamble por Gradient Boosting tienen mejor rendimiento que otros modelos, mejorando la precisión de los modelos. De hecho, se tomarán las experimentaciones de estos artículos para implementar nuestras comparaciones entre modelos, y decidiremos cual de estos modelos tendrá una mejor precisión.

III. Diseño del Experimento

A. Descripción del conjunto de datos

El conjunto de datos *in-vehicle coupon recommendation* [2] proporcionado por el repositorio de Aprendizaje Automático de la Universidad de California Irvine (UCI) contiene 12,684 registros etiquetados de forma binaria por la variable objetivo 'Y' con valores 0 y 1, siendo 1 el valor que indica cuando el potencial cliente "canjeará el cupón inmediatamente" o "lo cambiará más tarde pero antes de la fecha de vencimiento" y el valor 0 cuando el potencial cliente "no está interesado en recibir o intercambiar cupones". El número de muestras por clase en el conjunto de datos está distribuido de la siguiente manera: 7,210 muestras corresponden a la clase 1 y 5,474 muestras a la clase 2.

Asimismo, el conjunto de datos está compuesto por 26 características de los cuales 18 son tipo categórico y 8 características son de tipo numérico de los cuales 7 son binarios tomando valores 0 y 1. Existe una característica llamado "car" con un porcentaje de 99% de valores nulos y otras 4 características que apenas llegan al 1% de valores nulos como CoffeeHouse, CarryAway, RestaurantLessThan20 y Restaurant20To50.

Para nuestra experimentación tomaremos el 20% de la muestra total de 12,684 como conjunto de prueba, es decir contaremos con 10,147 muestras como conjunto de datos para realizar el entrenamiento y 2,537 muestras como conjunto de datos de validación.

B. Metodología

1) Pre-Procesamiento de datos: Al analizar el conjunto de datos notamos que existen valores de tipo categórico y numérico por ello que para hacer posible nuestra experimentación haremos la transformación de datos categóricos a numéricos indexando los valores enteros asignados un diccionario que permita luego restablecer los valores. Asimismo, del conjunto de datos notamos que existen valores faltantes, para ello reemplazaremos esos valores con la mediana de cada variable.

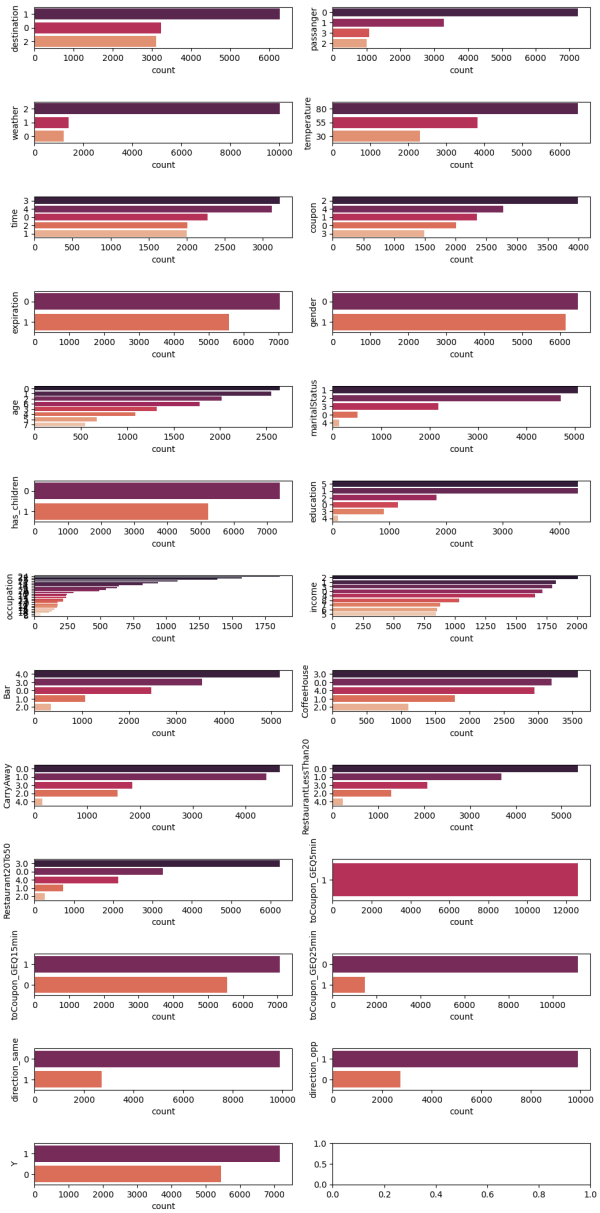

Fig. 1. Histograma de las variables del conjunto de datos.

Además, eliminaremos la variable (car) que tiene el 90% de valores nulos, también eliminaremos las variables que tienen redundancia por la similitud de sus valores como las variables *directionsame* y *directionopp*. Asimismo, se tiene una variable

de tipo numérico con un solo valor (toCoupon_GEQ5min), con correlación nula con respecto a las demás variables como se muestra en la Figura 1 y la Figura 2, la cual también se elimina.

Finalmente, para nuestra experimentación normalizaremos los datos utilizando estandarización estándar.

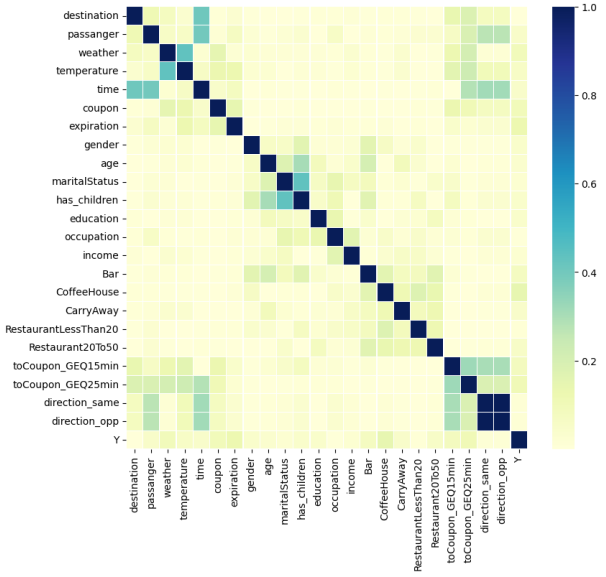

Fig. 2. Mapa de calor de la correlación entre variables del conjunto de datos

2) Selección y extracción de características: Para nuestra experimentación realizaremos una selección de características para encontrar las variables que son más importantes o tienen mayor impacto en el modelo de clasificación. La selección de características la realizaremos usando el método de filtrado aplicando coeficientes de correlación.

3) Selección y justificación de la medida de calidad: En vista de que nos hallamos ante una tarea de clasificación, y de que nos preocupan más una alta exactitud media sobre los errores de mayor magnitud, emplearemos como medida de calidad la exactitud (accuracy).

4) Algoritmos y estrategias de ajuste: Los algoritmos que emplearemos para la experimentación son los siguientes:

1) Árbol de Decisión (CART – Clasificación)
2) Clasificación Nearest Neighbor (KNN)
3) Random Forest – Clasificación
4) Extreme Gradient Boosting (XGBosst)
5) Clasificación con Regresión Logística
6) Support Vector Machine (SVM)

Para el entrenamiento de los modelos realizaremos regularización y optimización de hiperparámetros, además aplicaremos Validación Cruzada (CV) para evitar el sobreajuste del modelo.

5) Estrategia de validación: Realizaremos el ajuste de hiperparámetros a través de una búsqueda aleatoria para el descubrimiento de valores óptimos de nuestro modelo, usaremos las funciones GridSearchCV y RandomizedSearchCV propocionada por Scikit Learn. Asimismo, durante la experimentación aplicaremos ajuste manual de hiperparámetros en busca de mejores resultados.

REFERENCES

[1] A. S. Raamkumar and Y. Yang, Empathetic Conversational Systems: A Review of Current Advances, Gaps, and Opportunities. arXiv, 2022. doi: 10.48550/ARXIV.2206.05017.

[2] Wang, Tong, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry MacNeille. 'A bayesian framework for learning rule sets for interpretable classification.' The Journal of Machine Learning Research 18, no. 1 (2017): 2357-2393.

[3] P. Boteju and L. Munasinghe, "Vehicle Recommendation System using Hybrid Recommender Algorithm and Natural Language Processing Approach," in 2020 2nd International Conference on Advancements in Computing (ICAC), 2020, vol. 1, pp. 386–391. doi: 10.1109/ICAC51239.2020.9357156.

[4] Tran Duc Quynh and Hoang Thi Thuy Dung. "Prediction of Customer Behavior using Machine Learning: A Case Study". http://ceur-ws.org/Vol-3026/paper18.pdf (accedido may. 2022).

[5] E. Celik and S. Omurca, "Comparative Analysis of Offline Recommendation Systems with Machine Learning Algorithms," Jun. 2021.

[6] D. R. Hermawan, M. Fahrio Ghanial Fatihah, L. Kurniawati, and A. Helen, "Comparative Study of J48 Decision Tree Classification Algorithm, Random Tree, and Random Forest on In-Vehicle Coupon-Recommendation Data," in 2021 International Conference on Artificial Intelligence and Big Data Analytics, 2021, pp. 1–6. doi: 10.1109/ICAIBDA53487.2021.968970.

[7] P. Pokhrel and A. Lazar, "Towards Machine Learning Interpretability for Tabular Data with Mixed Data Types", FLAIRS, vol. 35, May 2022.