Un dibujo de una persona

Descripción generada automáticamente con confianza baja

INFORME DESARROLLO CASO DE ESTUDIO MÓDULO 2 - APRENDIZAJE SUPERVISADO

ANÁLITICA PARA LA TOMA DE DECISIONES

OSIRIS CONTRERAS TRILLOS

DAVID STEEVEN TAMAYO TORO

JUAN JOSÉ MOLINA OCAMPO

MARITZA ZAPATA GONZÁLEZ

DOCENTE

MANUELA LONDOÑO OCAMPO

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

2024 - S1

INTRODUCCIÓN

En un entorno cada vez más digitalizado, la inversión en marketing digital es fundamental para el éxito de las empresas que operan en línea. En este contexto, E-Corp, una empresa de venta de productos de lujo que incursionó hace un año en comercio electrónico busca mejorar la efectividad de su inversión en pauta digital para aumentar las conversiones y maximizar el retorno de la inversión.

Para lograr este objetivo, la dirección ha contratado a su equipo de consultores para abordar el problema descrito a través del uso de modelos de ML. En otras palabras, quieren conocer si pueden anticiparse a la identificación de clientes potenciales y así optimizar su inversión en pauta digital utilizando atributos que describen mejor a estos clientes potenciales.

METODOLOGÍA

Para el desarrollo del caso de estudio planteado, se trabajó con la base de datos ecommerce-data que corresponde al conjunto de datos de información relevante de las sesiones pertenecientes a usuarios diferentes en un período de 1 año y un diccionario de variables que contiene terminología de las variables para un mayor entendimiento de los datos.

Se llevó a cabo un análisis exploratorio de los datos para comprender su estructura y características. Se identificaron variables importantes que podrían influir en la probabilidad de compra de un visitante, como la duración de la visita, el número de páginas vistas y la interacción con contenido específico. Además, se realizó la limpieza de datos considerando los duplicados donde inicialmente se tenían 12330 filas y 18 columnas, se encontraron 125 duplicados que fueron eliminados. Posteriormente se convirtió el tipo de cada variable según la necesidad, se realizó la normalización y estandarización de variables para convertirlas en formato float y tener una escala aceptable, este último conjunto de datos de datos fue el insumo para construir los modelos.

La construcción de los modelos se hizo considerando la siguiente partición:

Tamaño del conjunto de entrenamiento: (9764, 17),

Tamaño del conjunto de validación: (2441, 17)

**Modelado de machine learning**

Se utilizaron técnicas de aprendizaje supervisado, como Regresión Logística, Árboles de Decisión y Random Forest, para construir modelos predictivos que identifiquen a los clientes potenciales.

En cada modelo se evaluó el rendimiento utilizando métricas de clasificación como precisión, recall y F1-score. Para optimizar la inversión en pauta digital, se pueden utilizar las predicciones del modelo para dirigir las campañas de marketing a los segmentos de clientes identificados como potenciales compradores.

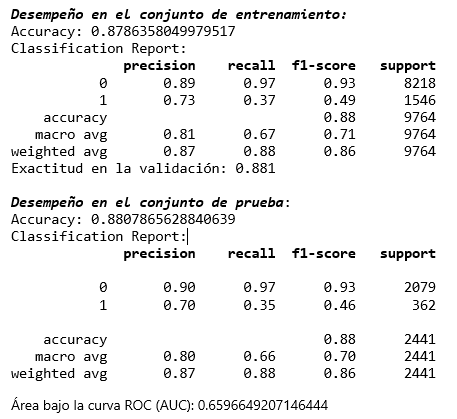
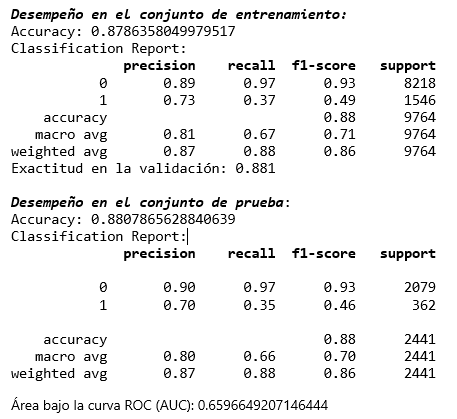
Para la iteración y mejora continua, este enfoque integral combina análisis de datos, modelado de machine learning y optimización estratégica para abordar el desafío específico de E-Corp y mejorar su rendimiento en el canal digital.

MODELOS CONSTRUIDOS Y RESULTADOS

Se construyeron 5 modelos de machine learning para predecir la decisión de compra de los visitantes del sitio web. A continuación, los resultados y métricas mencionadas anteriormente para cada uno de los modelos.

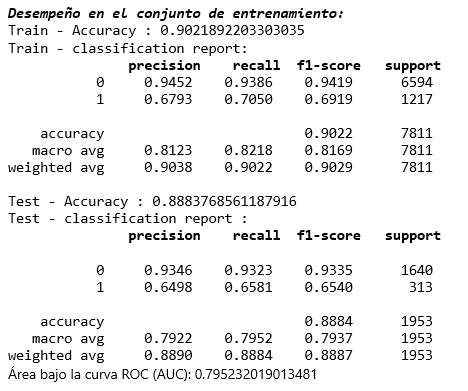
**Modelo 1: Regresión Logística**

Este modelo tiene un buen desempeño general, pero tiene dificultades para identificar correctamente las instancias de la clase positiva (compra), ya que su recall es relativamente bajo en comparación con la precisión. Esto puede indicar que el modelo necesita mejorar su capacidad para detectar correctamente las instancias de la clase positiva para ser más efectivo en este escenario.



**Modelo 2: Decision Tree Classifier**

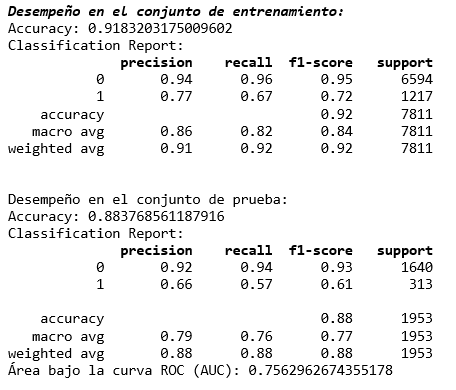
Este modelo parece generalizar bien a datos no vistos, pero hay un ligero desbalance de clases, lo que puede afectar las métricas de precisión y recall para la clase minoritaria (clase 1). Es importante tener en cuenta estas métricas al interpretar la efectividad del modelo y considerar ajustes según sea necesario.

 Tabla

Descripción generada automáticamente

**Modelo 3: Gradient Boosting**

Este modelo muestra un buen desempeño en términos de precisión y recall para ambas clases, aunque el F1-score para la clase 1 podría ser mejorado, este modelo podría beneficiarse a través de ajustes adicionales de hiperparámetros para mejorar su rendimiento.

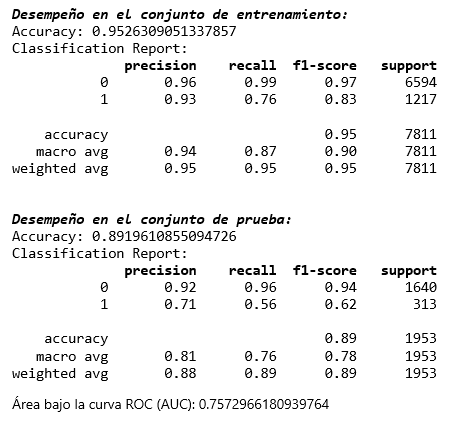
 Imagen de la pantalla de un celular con letras

Descripción generada automáticamente con confianza media

**Modelo 4: Random Forest**

Este modelo muestra un excelente desempeño en el conjunto de entrenamiento, con una alta precisión y recall para ambas clases. Además, generaliza bien para nuevos datos en el conjunto de prueba, con una precisión del 89.2%. Sin embargo, aunque su desempeño es bueno, se observa una ligera disminución en la precisión y recall para la clase minoritaria ("compran"), lo que sugiere que el modelo podría mejorar en la identificación de estas instancias. A pesar de esto, sigue siendo una opción sólida y competitiva para este.

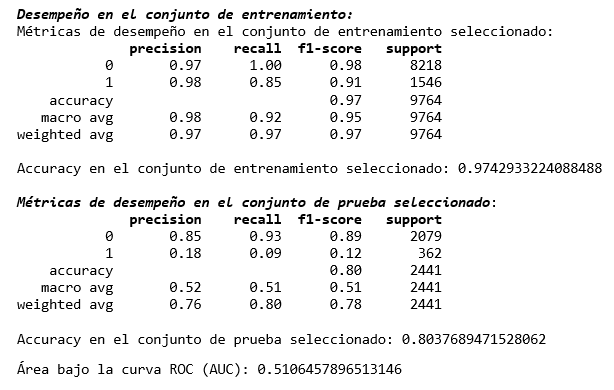
Imagen que contiene texto, recibo

Descripción generada automáticamente 

**Modelo 5: Random Forest con Lasso**

Con este modelo se utilizó el estimador Lasso, muestra un excelente desempeño en el conjunto de entrenamiento, con una alta precisión y recall para ambas clases. Sin embargo, en el conjunto de prueba, el AUC es notablemente bajo con respecto a los demás modelos.

Imagen de la pantalla de un celular de un mensaje en letras negras

Descripción generada automáticamente con confianza baja

MEJOR MODELO

Si nos basamos en la exactitud en el conjunto de prueba, el Modelo 2 (DecisionTreeClassifier) y el Modelo 4 (Random Forest Classifier) son los mejores, con una exactitud del 89%.

El F1-score es la media armónica de precisión y recall. Es útil cuando hay una clase desbalanceada. En este caso para el modelo 2, el F1-score para la clase 0 es de 93%, y para la clase 1 es de 65% en la validación. Mientras que en el modelo 4 para la clase 0 es de 94% y para la clase 1 es de 62% también en los datos de validación.

Sin embargo, también es crucial considerar otras métricas como el AUC. Dado que éste mide la capacidad de discriminación del modelo, que es importante especialmente en problemas de clasificación desbalanceada como este.

El Modelo 2 (DecisionTreeClassifier) se destaca como el mejor debido a su equilibrio entre una alta exactitud del 89% y la mayor AUC de 0.80 en el conjunto de prueba, lo que indica una fuerte capacidad de discriminación, especialmente valiosa en contextos de clasificación desbalanceada como es nuestro caso.

CONCLUSIONES Y RECOMENDACIONES

**1. Estrategias de Marketing:** Basado en los resultados de los modelos, hay margen de mejora en las estrategias de marketing digital. Los datos sugieren que la nueva estrategia de marketing, implementando el modelo 2 podría ser más efectiva ya que tiene un enfoque en la personalización y la segmentación para captar nuevos clientes con alta intención de compra.

**2. Utilización de Modelos de Machine Learning:** Los modelos de ML en este caso son valiosos para identificar clientes potenciales. El Modelo 2, en particular, es prometedor en la identificación de estos clientes, optimizando así la inversión en publicidad digital. Este modelo destaca por su capacidad predictiva y debería ser el foco de futuras inversiones y desarrollos.

**3. Refinamiento del Modelo:** Aunque el Modelo 2 parece ser el más efectivo hasta ahora, presenta oportunidades de mejora. Se pueden explorar otras técnicas de ajuste de hiperparámetros y evaluar la implementación de algoritmos alternativos para maximizar la precisión y el retorno de la inversión.

**4. Monitoreo Continuo:** El dinamismo del mercado requiere una adaptación constante dado que las preferencias de los clientes pueden cambiar con el tiempo, esto implicaría la toma de datos y nuevas variables que puedan incidir en la compra del cliente. Es crucial implementar un sistema de monitoreo que permita la actualización continua del modelo en respuesta a las nuevas tendencias de consumo y cambios en el entorno de mercado.

En conclusión, el Modelo 2 (DecisionTreeClassifier) es el mejor modelo debido a su alta exactitud y AUC en el conjunto de prueba. Sin embargo, la elección del mejor modelo puede variar según la importancia relativa de cada métrica y las características específicas del problema.