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 Lineal, Árboles de Decisión y Tunning de hiperparámetros, para construir modelos predictivos que identifiquen a los clientes potenciales.

MODELOS CONSTRUIDOS Y RESULTADOS

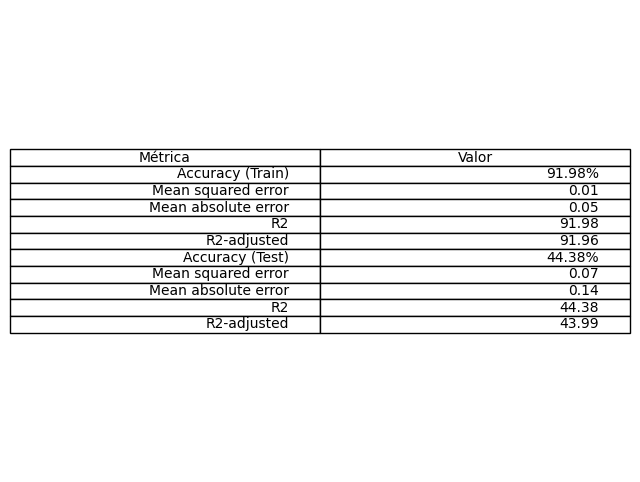
Se construyeron varios modelos de machine learning para predecir la decisión de compra de los visitantes del sitio web.

Cada modelo se ajustó y evaluó utilizando métricas de desempeño como precisión, error cuadrático medio (MSE), error absoluto medio (MAE) y coeficiente de determinación (R2).

**Modelo 1: Random Forest sin selección de variables**

Se utilizó el algoritmo Random Forest para predecir la decisión de compra de un visitante sin realizar una selección previa de variables y las métricas para este modelo fueron las siguientes:

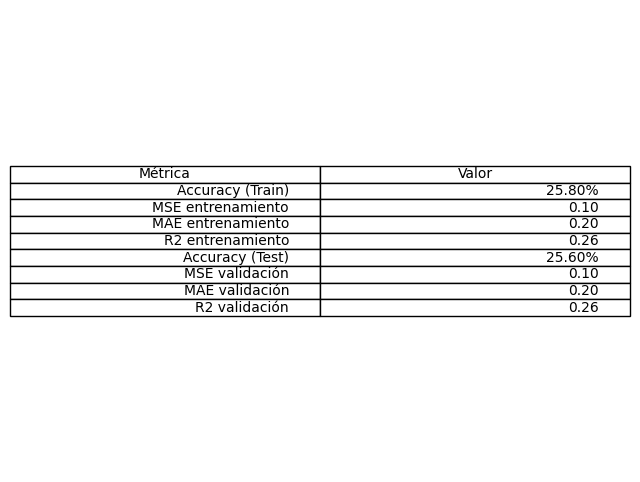
Se observó que este modelo proporcionaba una alta precisión en el conjunto de entrenamiento, pero una precisión menor en el conjunto de validación.



**Modelo 2: Regresión Lineal Múltiple con selección de variables**

Se implementó una regresión lineal múltiple utilizando la técnica Lasso para seleccionar las variables más relevantes.

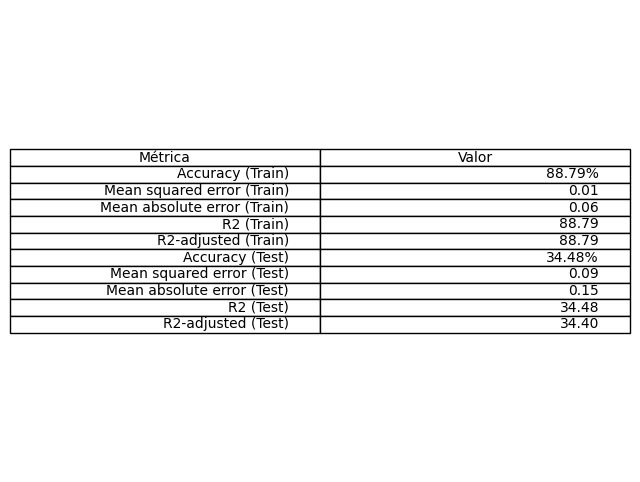
Se identificaron variables importantes como 'Reviews\_Duration', 'ProductRelated\_Duration' y 'PageValues' que influyen significativamente en la probabilidad de compra y se obtuvieron las siguientes métricas:



**Modelo 3: Random Forest con selección de variables**

Se utilizó Random Forest nuevamente, pero esta vez se realizaron selecciones de variables, eliminando las variables con coeficiente cero.

Se encontró que este modelo proporcionaba una precisión comparable al Modelo 1 pero con un número reducido de variables, lo que sugiere una mejor generalización. Las métricas del modelo se presentan a continuación:

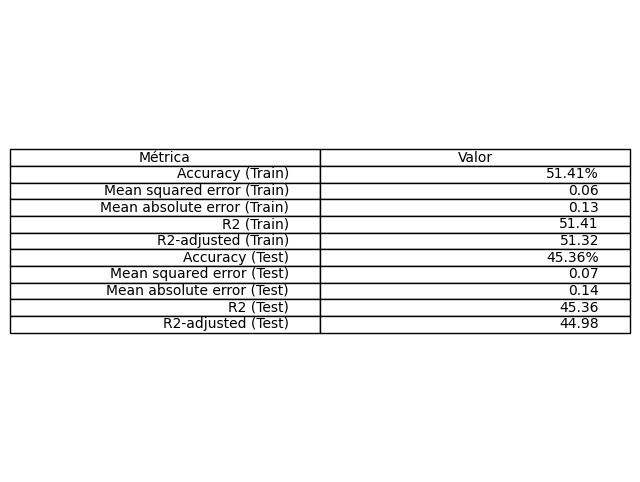


**Modelo 4: Tuning de Hiperparâmetros para Random Forest**

Se llevó a cabo una búsqueda aleatoria de hiperparámetros para optimizar el desempeño del modelo Random Forest.

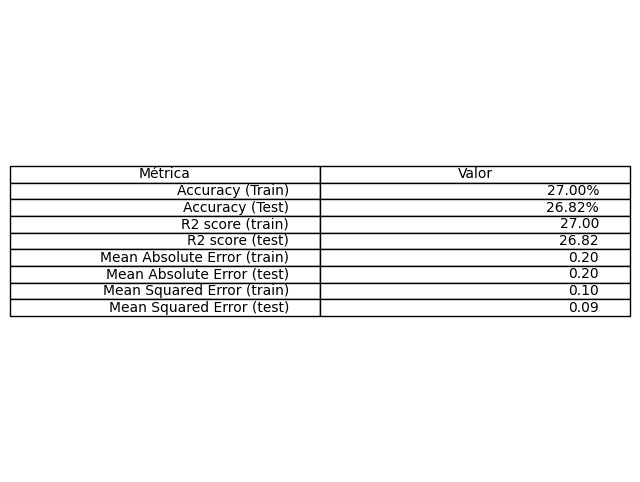
Se observó una ligera mejora en la precisión del modelo con los datos de testeo y desmejoró en el entrenamiento después de ajustar los hiperparámetros. Las métricas fueron las siguientes:

{'n\_estimators': 150, 'min\_samples\_split': 5, 'min\_samples\_leaf': 4, 'max\_depth': 5, 'bootstrap': True}



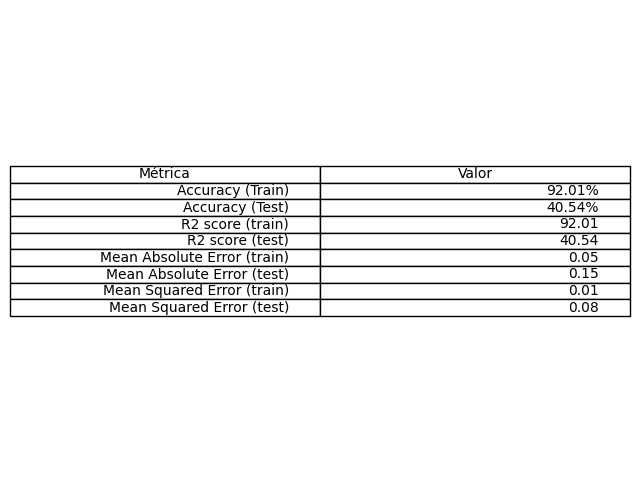
**Modelo 5: Regresión lineal múltiples con características Variance Threshold**

Se aplicó una regresión lineal múltiple utilizando la técnica de Variance Threshold para seleccionar características, con el respecto al modelo 2 tiene una leve mejora, sigue teniendo dificultades para predecir con precisión los valores de la variable objetivo. Las métricas fueron las siguientes:

****

**Modelo 6: Random Forest con características Variance Threshold**

Se aplicó el Random Forest utilizando la técnica de Variance Threshold para seleccionar características, logra una alta precisión en el conjunto de entrenamiento (92.008%), pero su desempeño en el conjunto de prueba (40.543%) es relativamente inferior, lo que indica cierto sobreajuste. Las métricas fueron las siguientes:

****

MEJOR MODELO

Modelo 6: **Random Forest con características Variance Threshold:**

* Accuracy (Train): 92.01%
* Mean squared error (Train): 0.01
* Mean absolute error (Train): 0.05
* R2 (Train): 92.01%
* Accuracy (Test): 40.54%
* Mean squared error (Validation): 0.08
* Mean absolute error (Validation): 0.15
* R2 (Validation): 40.54%

El Modelo 1 sobresale en capacidad predictiva general con un alto (R^2), mientras que el Modelo 6, a pesar de un (R^2) más bajo, podría ser preferible para identificar clientes potenciales y optimizar la inversión en publicidad digital. La elección entre ambos depende de si se prioriza la precisión predictiva o la identificación de clientes.

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 4 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 6, 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 6 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.