Informe Caso de Estudio – Aprendizaje Supervisado

Integrantes:

Natalia Ximena Guzmán Zorrilla

Sebastián Cuartas Gil

Alexander Amaya León

Universidad de Antioquia, Facultad de Ingeniería

Analítica para la Toma de Decisiones [2502072]

Manuela Londoño Ocampo

Medellín, Colombia

01 de abril de 2024

Exploración de los Datos

Inicialmente, se realiza una exploración detallada de los datos disponibles, los cuales contienen información sobre las interacciones de los usuarios en el sitio web de E-Corp. Se examinan diversas variables, tanto numéricas como categóricas, para comprender su distribución y características.

Análisis Univariado y Bivariado

Se lleva a cabo un análisis univariado, donde se presentan histogramas de las variables numéricas y gráficos de conteo para las variables categóricas para así determinar el comportamiento de los datos. Además, se realiza un análisis bivariado para explorar la relación entre las variables numéricas y la variable objetivo "Purchase", donde por medio de diagramas de boxplot comparativo, se puede apreciar la gran cantidad de valores extremos que hay en la mayoría de las variables. Sin embargo, se opta por un proceso de winzorizado para mitigar un poco la influencia que puedan tener estos valores en los modelos. También se investigan las relaciones entre las variables, tanto numéricas como categóricas, para identificar posibles relaciones. Además, por medio de la matriz de correlación nos dimos cuenta que existen variables explicativas que están correlacionadas entre sí, por lo que se procede a eliminarlas para evitar problemas de multicolinealidad.

Selección de Variables

Para la selección de variables, primero se eliminaron características que tuviesen relaciones entre ellas (todas las variables de tiempo de duración), después se utilizaron metodologías integradas y wrapper para reducir la complejidad del dataset y evitar posibles sobreajustes. Como resultado, se obtuvo que hay variables más significativas que otras, pasando de tener 73 variables a tener 30 por método de wrapper y 20 por método integrado.

Modelos Utilizados

Los modelos que se utilizaron para predecir sobre los datos fueron la regresión logística tradicional y los modelos de ensamble como los random forest y el XGB classifier porque el problema era un problema de clasificación.

Resultados obtenidos

	Train						Test								
	Accuracy	Precisión		Recall		F1	Accuracy	Precisión		Recall		F1	AUC	MC	
		0	1	0	1			0	1	0	1				
Modelo base	0.8746	0.8804	0.7815	0.9848	0.2891	0.6759	0.8824	0.8869	0.8	0.988	0.2762	0.6727	0.8833	2054 262	25 100
Modelo con balanceo	0.8181	0.9521	0.4564	0.8254	0.7794	0.73	0.8267	0.9515	0.4498	0.8393	0.7541	0.7277	0.8926	1745 89	334 273
Regresión Logística con submuestreo	0.7801	0.9529	0.3918	0.7788	0.7874	0.6902	0.7832	0.9475	0.4235	0.7827	0.7864	0.7039	0.8626	1588 88	441 324
Regresión Logística con sobremuestreo	0.828	0.9545	0.4636	0.8368	0.7796	0.7366	0.813	0.9452	0.4515	0.825	0.7506	0.7224	0.8861	1707 99	362 298
Regresión Logística con método wrapper	0.813	0.95	0.44	0.82	0.78	0.72	0.817	0.94	0.47	0.83	0.75	0.73	0.789	1710 104	345 307
Regresión Logística con método integrado	0.816	0.95	0.44	0.82	0.78	0.72	0.819	0.74	0.47	0.83	0.75	0.73	0.789	1711 103	344 308
Random Forest con método wrapper	0.9998	0.9999	1	1	0.9994	0.9998	0.9061	0.9267	0.7436	0.9663	0.5608	0.7927	0.9111	2009 159	70 203
Random Forest con método integrado	0.9995	0.9998	0.9987	0.9998	0.9987	0.9992	0.902	0.9279	0.7113	0.9596	0.5718	0.7887	0.9196	1995 155	84 207
Random Forest integrado con tuneo de hpp	0.8667	0.9695	0.5511	0.8691	0.8545	0.7933	0.8607	0.9652	0.5192	0.8677	0.8204	0.7749	0.9291	1804 65	275 297
Random Forest wrapper con tuneo de hpp	0.8647	0.971	0.546	0.8651	0.8629	0.7919	0.8603	0.9662	0.5182	0.8663	0.826	0.7752	0.9218	1801 63	278 299
XGB wrapper	0.9224	0.9423	0.7968	0.9671	0.685	0.8456	0.907	0.9331	0.7228	0.9596	0.605	0.8024	0.9239	1995 143	84 219
XGB integrado	0.9177	0.9402	0.777	0.9636	0.674	0.8368	0.909	0.9337	0.7333	0.9615	0.6077	0.806	0.9322	1999 142	80 220
XGB integrado con tuneo de hpp	0.9209	0.9415	0.7905	0.9661	0.6811	0.8427	0.9045	0.9313	0.7143	0.9586	0.5939	0.7967	0.9342	1993 147	86 215
XGB wrapper con tuneo de hpp	0.9402	0.9485	0.8843	0.9824	0.7167	0.8784	0.9029	0.9244	0.7306	0.9649	0.547	0.7849	0.9304	2006 164	73 198

Tabla 1. Métricas de los modelos.

Comparativa entre modelos base y con balanceo

Al comparar los modelos, el modelo con balanceo mostró un recall significativamente más alto para la clase positiva tanto en el conjunto de entrenamiento (0.7794) como en el conjunto de prueba (0.7541), en contraste con el modelo base que registró valores de 0.2891 y 0.2762 respectivamente. Además, el AUC del modelo con balanceo fue ligeramente superior (0.8926) en comparación con el modelo base (0.8833). Estas métricas indican una mejor capacidad del modelo con balanceo para detectar casos positivos y una mejor capacidad de discriminación entre clases, lo que lo posiciona como el modelo preferido en esta evaluación.

Comparativa entre modelos de regresión logística

Al comparar los modelos de regresión logística que emplean diferentes técnicas de muestreo y selección de características, se destacan varias diferencias significativas en su desempeño. El modelo de regresión logística con sobremuestreo se distingue por presentar el recall más alto para la clase positiva en el conjunto de prueba, con un valor de 0.7864. Esto sugiere una mejor capacidad para detectar casos positivos, lo cual es importante en un escenario donde la identificación de clientes potenciales es fundamental. Pero, este modelo exhibe el F1-score menos alto (macro avg) en el conjunto de prueba, alcanzando 0.7039, lo que indica un buen equilibrio entre precisión y recall. También, se destaca que el AUC más alto entre los modelos comparados es el del modelo con sobremuestreo, con un valor de 0.8861, lo que refleja un rendimiento general superior en la clasificación binaria y una mayor capacidad para distinguir entre las clases. En contraste, el modelo de regresión logística con submuestreo muestra un recall ligeramente más alto que el modelo con sobremuestreo, pero presenta el AUC más bajo. Por otro lado, los modelos con método wrapper y método integrado muestran resultados muy similares en todas las métricas evaluadas, con ligeras variaciones insignificantes. Aunque presentan un rendimiento aceptable, no logran superar al modelo con sobremuestreo en términos de recall y AUC. Por lo tanto, considerando el contexto del caso de estudio y los objetivos comerciales, se puede concluir que el modelo de regresión logística con sobremuestreo es el más equilibrado y efectivo para anticiparse a la identificación de clientes potenciales en el sitio web de E-Corp.

Comparativas entre modelos Random Forest

Al analizar los modelos de Random Forest con diferentes técnicas de selección de características y ajuste de hiperparámetros, se observan algunas diferencias importantes en su desempeño. Los modelos Random Forest con método wrapper y método integrado muestran un sobreajuste evidente, con un rendimiento perfecto (o casi perfecto) en el conjunto de entrenamiento, pero un rendimiento inferior en el conjunto de prueba, indicado por los valores de 1 en varias métricas, como precisión, recall y F1-score. Esto sugiere que estos modelos pueden haber capturado demasiado el ruido en los datos de entrenamiento y no generalizan bien a datos no vistos.

En contraste, los modelos Random Forest con ajuste de hiperparámetros muestran un rendimiento más equilibrado en el conjunto de prueba, con valores de precisión, recall y F1-score similares al conjunto de entrenamiento, pero aún bastante altos, lo que indica un menor sobreajuste. Específicamente, el modelo Random Forest integrado con ajuste de

hiperparámetros muestra el AUC ligeramente más alto entre los modelos evaluados, con un valor de 0.9291, lo que sugiere una mejor capacidad para distinguir entre las clases en el conjunto de prueba. Sin embargo, los modelos Random Forest con método wrapper y método integrado también muestran un AUC bastante alto, con valores de 0.9111 y 0.9196 respectivamente, lo que indica un rendimiento competitivo en la clasificación binaria, pero teniendo en cuenta que son modelos que están sobreajustados.

Considerando el contexto del caso de estudio de E-Corp, donde se busca identificar clientes potenciales para optimizar la inversión en publicidad digital, se puede concluir que el modelo Random Forest integrado con ajuste de hiperparámetros es el más equilibrado y efectivo con métricas de recall como 0.8677 (0) y 0.8204 (1), al igual que el modelo de Random Forest con método wrapper y ajuste de hiperparámetros con valores similares en las métricas y el AUC, lo que los hace más adecuados para cumplir con los objetivos de negocio de E-Corp.

Comparativas entre modelos XGBoost

Al evaluar los modelos XGBoost aplicados con diferentes enfoques de selección de características y ajuste de hiperparámetros, es importante destacar que, si bien muestran un buen desempeño en general, el recall en la clase positiva no es significativamente alto en comparación con el conjunto total de datos. Es decir, aunque estos modelos son capaces de identificar clientes potenciales en cierta medida, no logran capturar una proporción considerable de los casos positivos, es decir, aquellos usuarios que realizan compras en el sitio web de E-Corp.

Este hallazgo sugiere que los modelos pueden estar pasando por alto una parte significativa de los clientes que generan ingresos para la compañía. En particular, el modelo XGBoost wrapper con ajuste de hiperparámetros muestra un recall del 0.96 tanto en el conjunto de entrenamiento como en el de prueba, lo que indica una capacidad consistente para identificar correctamente la mayoría de las transacciones de no compra en los datos utilizados para entrenar el modelo y en datos nuevos. Sin embargo, es importante destacar que el recall en la clase positiva no supera el 0.8, lo que sugiere que aún existe margen para mejorar la capacidad de los modelos para detectar clientes potenciales de manera más efectiva. Esta observación es relevante para E-Corp, ya que indica la necesidad de ajustar y mejorar continuamente las estrategias de marketing digital para maximizar el retorno de la inversión en publicidad y optimizar las ventas a través del canal en línea.

Selección del modelo

Para realizar un análisis global de todas las métricas proporcionadas y seleccionar dos modelos, se debe considerar una combinación de varias métricas clave, incluido el recall en la clase positiva, la precisión en la clase negativa, la precisión global y el AUC. Estos criterios son fundamentales para evaluar el rendimiento de los modelos en la identificación de clientes potenciales y la maximización de las ventas en el sitio web de E-Corp.

Tras revisar detenidamente las métricas de todos los modelos, se han seleccionado los siguientes dos modelos:

1. Random Forest integrado con ajuste de hiperparámetros: Este modelo muestra un recall muy equilibrado en el conjunto de entrenamiento y en el conjunto de prueba, lo que indica una capacidad consistente para identificar correctamente la mayoría de las transacciones de compra en los datos utilizados tanto para entrenar el modelo como para evaluarlo en datos nuevos. Además, presenta una alta precisión en la clase negativa (usuarios que no realizan compras). Esto sugiere que el modelo es efectivo para distinguir entre los usuarios que realizan compras y los que no, minimizando los falsos positivos. Además, el AUC indica un buen rendimiento global del modelo en la clasificación binaria.

2. Random Forest wrapper con ajuste de hiperparámetros: Este modelo también muestra un buen rendimiento en términos de recall en la clase positiva tanto en el conjunto de entrenamiento como en el de prueba. Aunque ligeramente inferior al modelo anterior, sigue siendo considerablemente alto y sugiere una capacidad razonable para identificar clientes potenciales. Además, presenta una alta precisión en la clase negativa en ambos conjuntos, lo que indica una capacidad para distinguir efectivamente entre usuarios que realizan compras y los que no. El AUC también indica un buen rendimiento global del modelo en la clasificación binaria.

Ambos modelos muestran un equilibrio sólido entre la capacidad para detectar clientes potenciales y la precisión en la identificación de usuarios que no realizan compras. Sin embargo, se han seleccionado estos modelos específicos debido a su buen rendimiento en el recall de la clase positiva, lo que sugiere una capacidad razonable para identificar clientes potenciales de manera efectiva. Esta selección se alinea con el objetivo principal del caso de estudio de E-Corp, que es optimizar las estrategias de marketing digital para maximizar las ventas a través del canal en línea y mejorar el retorno de la inversión en publicidad.

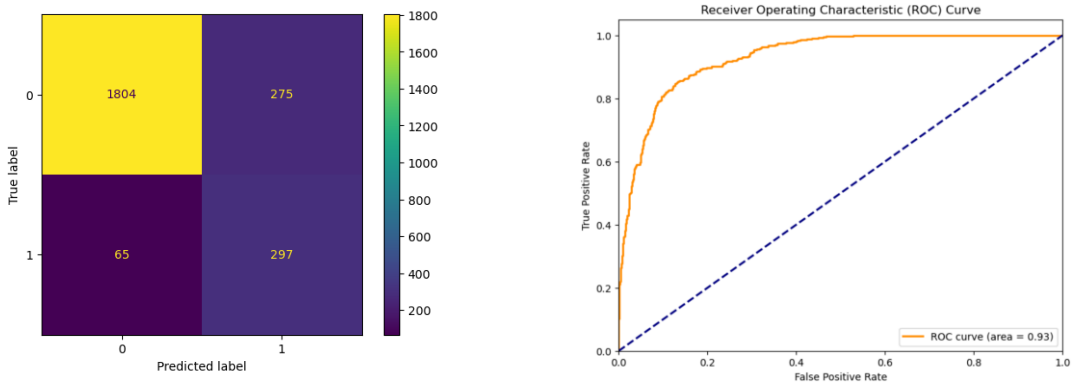

Figura 1. Gráficos del modelo Random Forest Integrado con tuneo de hiperparámetros.

La matriz de confusión muestra que el modelo tiene 297 verdaderos positivos (TP), 275 falsos positivos (FP), 1804 verdaderos negativos (TN) y 65 falsos negativos (FN). Estos valores indican la capacidad del modelo para predecir correctamente la compra y la no compra. Aunque hay un número considerable de predicciones correctas, los falsos positivos y falsos negativos sugieren que el modelo aún puede mejorar en su precisión. En el contexto del caso de estudio de E-Corp, estos resultados significan que el modelo puede identificar clientes potenciales,

pero también puede generar gastos innecesarios en campañas de marketing dirigidas a usuarios que no tienen la intención de comprar.

Conclusiones

Después de un exhaustivo análisis de los modelos implementados en el contexto del caso de estudio de E-Corp, podemos concluir lo siguiente:

1. Tras evaluar los múltiples modelos de aprendizaje automático, se identificó que los modelos basados en Random Forest y XGBoost integrados con ajuste de hiperparámetros ofrecen el mejor equilibrio entre rendimiento y generalización. Estos modelos destacan por su capacidad para capturar patrones complejos en los datos y su capacidad para adaptarse a diferentes conjuntos de datos sin caer en el sobreajuste.
2. A pesar de los resultados prometedores, los modelos aún enfrentan desafíos significativos en la predicción de la intención de compra de los usuarios en el sitio web de E-Corp. Se observa que el recall en la clase positiva no es significativamente alto, lo que sugiere que los modelos tienen dificultades para identificar correctamente a los usuarios que están propensos a realizar una compra. Esta debilidad puede atribuirse a la complejidad del comportamiento del usuario en línea y a la necesidad de características más sofisticadas para capturar adecuadamente la intención de compra.
3. A pesar de los desafíos identificados, los modelos implementados ofrecen una oportunidad prometedora para optimizar las campañas de marketing digital de E-Corp. Al identificar de manera más precisa a los usuarios potenciales y enfocar los esfuerzos de marketing en estos segmentos específicos, la compañía puede maximizar el retorno de su inversión publicitaria y mejorar la eficiencia de sus estrategias de adquisición de clientes.
4. Es fundamental reconocer que el desarrollo de modelos predictivos es un proceso continuo que requiere iteración y mejora constante. A medida que se acumulan más datos y se implementan nuevas estrategias, es necesario revisar y actualizar los modelos para mantener su relevancia y precisión. Además, se recomienda la exploración de nuevas técnicas de modelado y la incorporación de características adicionales para enriquecer la capacidad predictiva de los modelos.

Recomendaciones

1. Se recomienda realizar un análisis exhaustivo de las características utilizadas en los modelos, con el objetivo de identificar aquellas que contribuyan de manera más significativa a la predicción del comportamiento de los clientes en el sitio web. Este proceso puede incluir la incorporación de nuevas variables o la ingeniería de características para mejorar la capacidad predictiva de los modelos.
2. Se sugiere establecer un sistema de monitoreo continuo para evaluar el desempeño de los modelos en tiempo real y detectar posibles desviaciones o deterioros en su rendimiento. Además, es fundamental implementar un proceso de mejora iterativa que permita ajustar y actualizar los modelos en función de los cambios en los datos o en el comportamiento de los usuarios.