MAURO SALINAS

Predicción de clientes potenciales

Informe de resultados del modelo de predicción

Índice

- Introducción
- Hipótesis
- Data Wrangling
- Exploratory Data Analyses
- Modelo de Predicción
- Conclusiones

Introducción

Una compañía de seguros de autos cuenta con una base de datos de los distintos individuos a los que les realizó una cotización de seguro.

Este DataSet contiene características del individuo, características del contacto, variables asociadas al monto de la cotización y si finalmente el individuo contrató o no el servicio de seguros de autos.

Con el fin de concentrar sus esfuerzos sobre los clientes más probables a comprar y ahorrar recursos al no esforzarse en los menos probables, el objetivo de este análisis será predecir la variable Conversion_Status por medio de Aprendizaje Supervisado por Clasificación.

Variables

La variable objetivo es Conversion_Status. Es binaria tal que “1” indica que el cliente compró.
A continuación, el resto de las variables:

1. Age (numérico): Edad del sujeto.
2. Is_Senior (texto): Binario que indicado si la persona es o no mayor a 55 años.
3. Marital_Status (texto): Estado Civil.
4. Married_Premium_Discount (número): Descuento en caso de casado.
5. Prior_Insurance (texto): Duración del seguro anterior al actual.
6. Prior_Insurance_Premium_Adjustment (número): Descuento por duración del seguro anterior. 50 en caso de 1-5 años y 100 en caso de <1 año.
7. Claims_Frequency (número): frecuencia anual de reclamos.
8. Claims_Severity (texto): severidad de los reclamos.
9. Claims_Adjustment (número): ajuste por reclamos (Frecuencia por severidad, considerando la severidad Low = 50, Medium = 100 y High = 200).
10. Policy_Type (texto): tipo de cobertura (Full Coverage o Liability-Only).
11. Policy_Adjustment (número): ajuste por Policy_Type (si es Liability-Only = -200, Full Coverage = 0).
12. Premium_Amount (número): Promoción final.
13. Safe_Driver_Discount (texto): binario que indica con "1" si corresponde descuento por buen conductor.
14. Multi_Policy_Discount (texto): binario que indica con "1" si corresponde descuento por múltiples.
15. Bundling_Discount (texto): binario que indica con "1" si corresponde descuento por paquetes.
16. Total_Discounts (número): Indica el descuento total que corrresponde a las últimas 3 variables. Suma los últimos 3 y los multiplica por 50.
17. Source_of_Lead (texto): medio de contacto con el cliente. Puede ser Online, Agent o Referral.
18. Time_Since_First_Contact (número): tiempo en días desde el primer contacto.
19. Inquiries (número): cantidad de consultas.
20. Quotes_Requested (número): cantidad de cotizaciones solicitadas.
21. Time_to_Conversion (número): tiempo estimado para la conversión.
22. Credit_Score (número): puntaje por historial crediticio.
23. Premium_Adjustment_Credit (número): ajuste que depende del historial crediticio. Si el primer dígito de Credit_Score es =>7, entonces -50, sino 50.
24. Region (texto): tipo de región del conductor. Puede ser Rural, Suburban o Urban.
25. Premium_Adjustment_Region (número): ajuste por región siendo Rural = 0, Suburban = 50 y Urban = 100.

Hipótesis

Efecto de descuentos

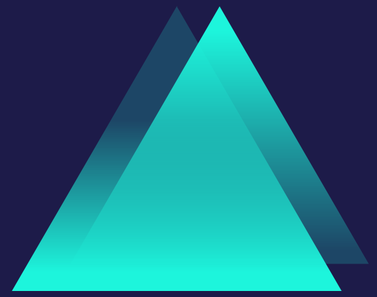
Los descuentos afectan el índice de conversión

Efecto del cliente

Las características del cliente están relacionadas al índice de conversión.

Efecto del producto

Los productos (Policy_Type) cuentan con distintos índices de conversión.



Efecto financiero

El historial crediticio del cliente impacta sobre el índice de conversión.

Data Wrangling

Limpieza de datos

0 %

Duplicados

No existen valores duplicados en los datos

0 %

Nulos

No existen valores nulos en los datos

100 %

Valores únicos eficaces

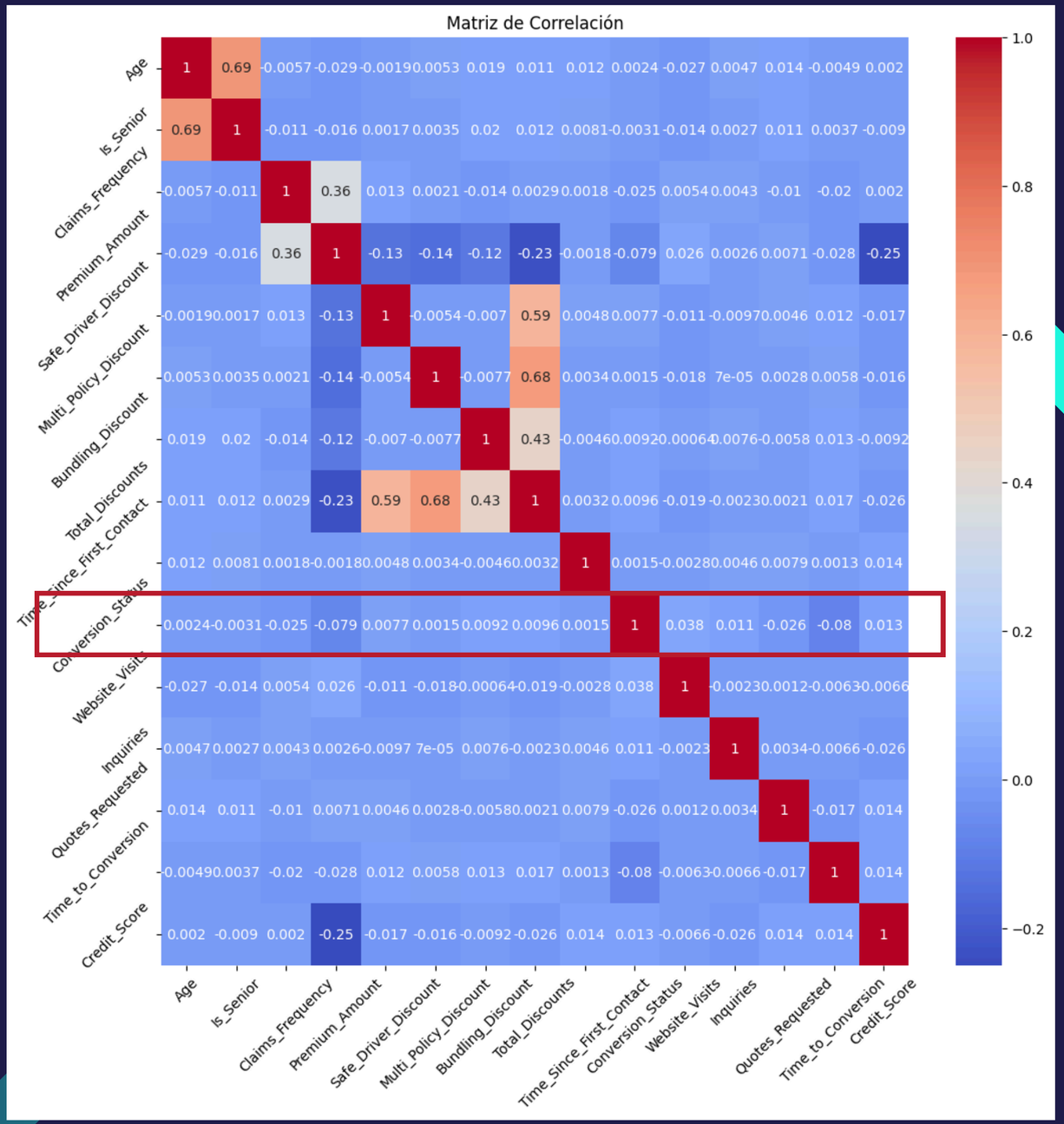
Los valores únicos son los apropiados. No se requiere estandarización de valores.

Exploratory Data Analyses

Bivariado

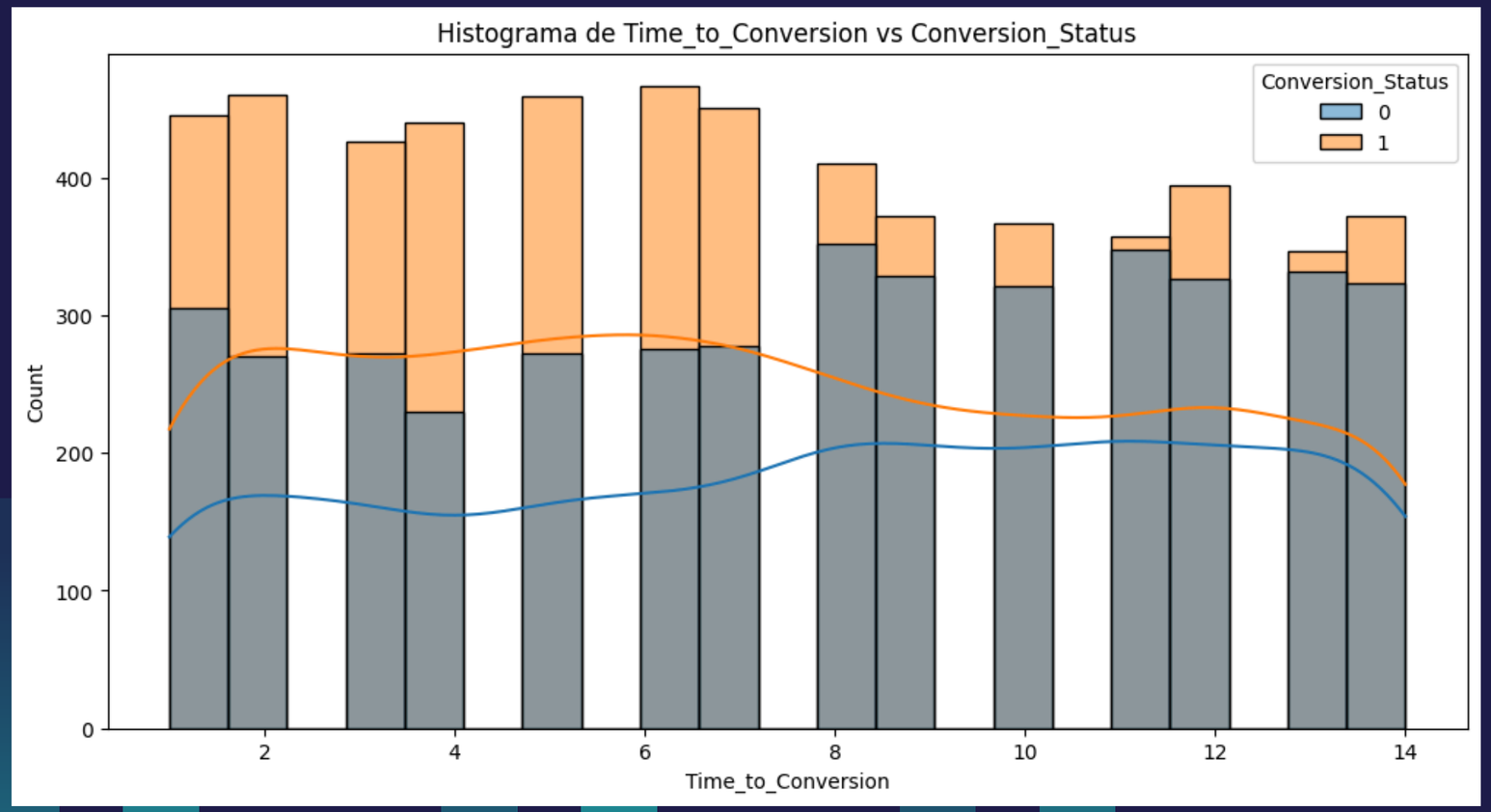
Variables numéricas

En primera instancia, no hay una correlación directa entre Conversion_Status y el resto de variables



Bivariado

Time_to_Conversion

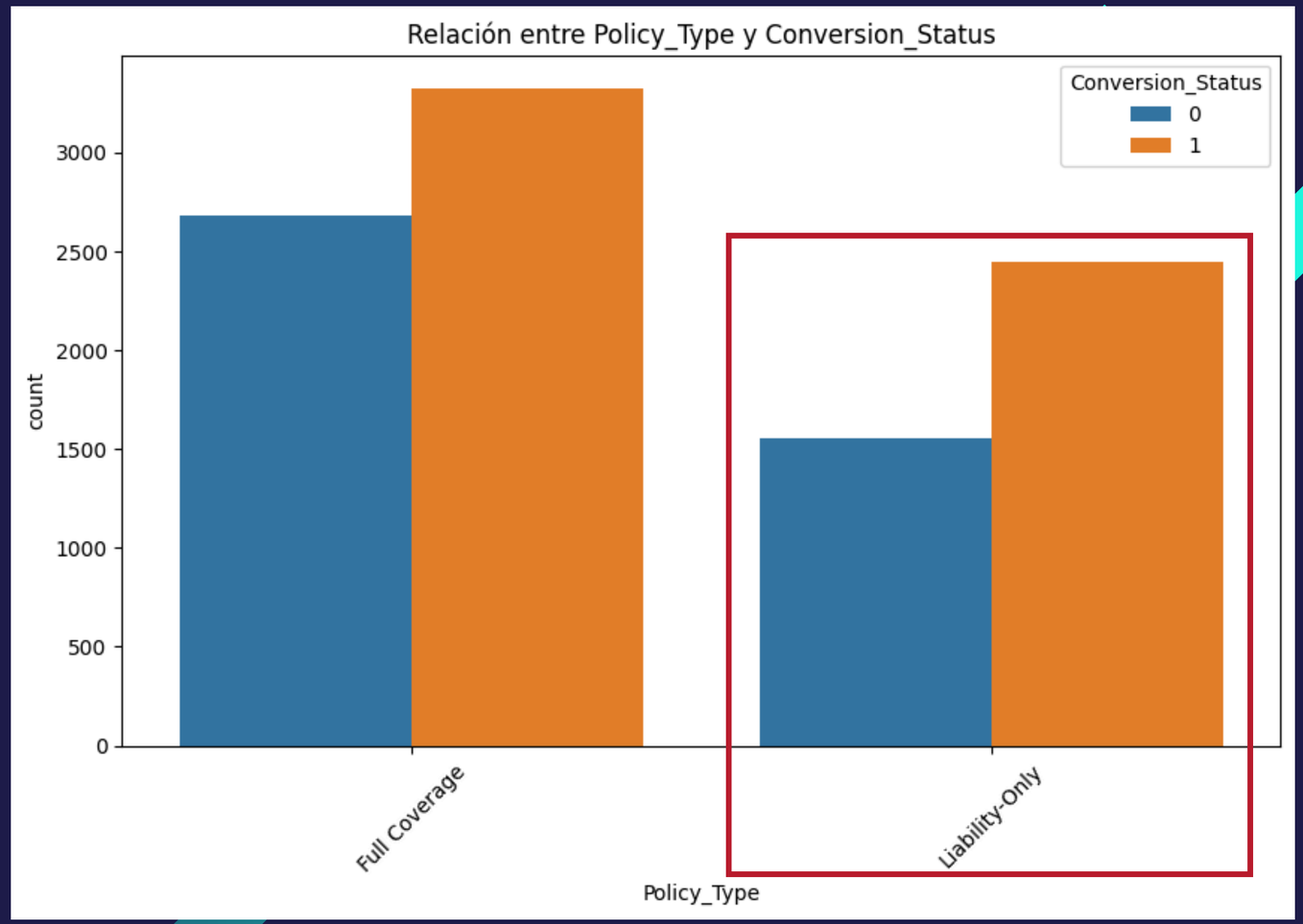


Hay una clara distinción en esta variable de que a valores bajos hay mayor tendencia a la compra.

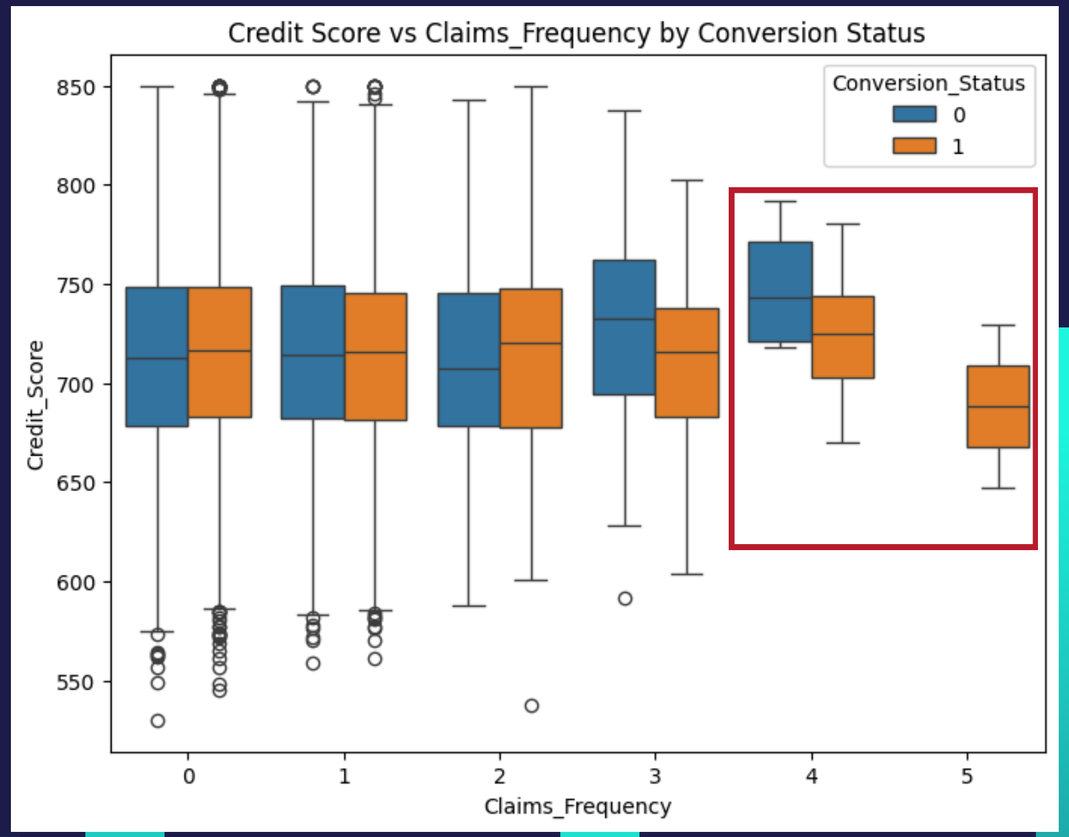
Bivariado

Variables categóricas

Se identifica en el caso de la variable Policy_Type que su valor “Liability-Only” presenta mayor tendencia a la compra.

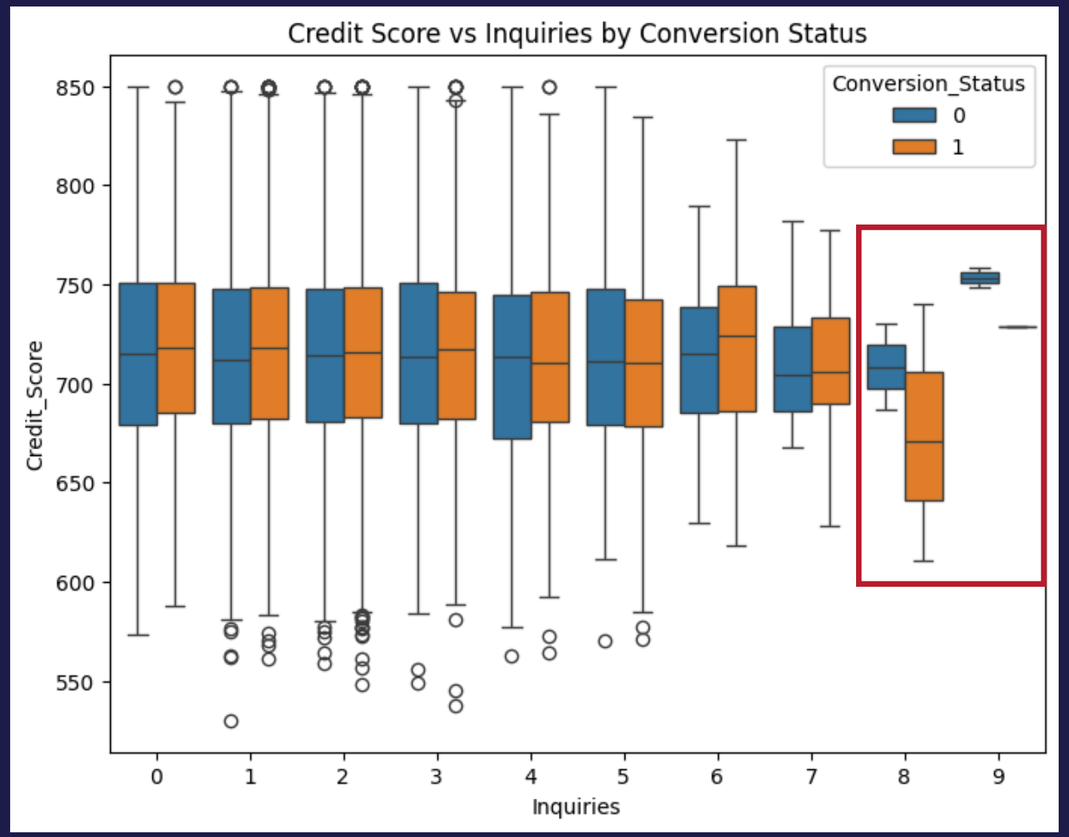


Multivariado



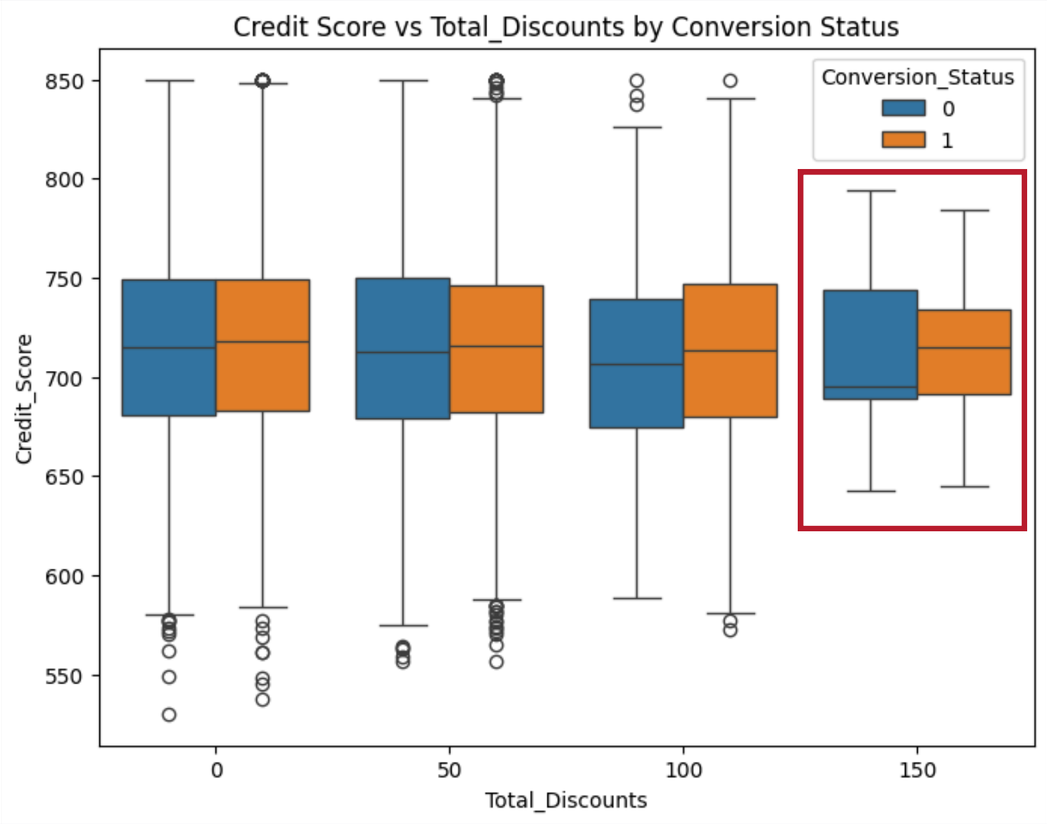
Al evaluar el Credit_Score junto a otras variables como Claims_Frequency o Inquiries, se logra distinguir un comportamiento distinto de los compradores positivos.

La lectura es que a valores altos de Claims_Frequency y Inquiries, los valores bajos de Credit_Score tienden a ser de compradores positivos.



Multivariado

Por otro lado, a valores altos de Total_Discounts, los compradores positivos tienden a tener Credit_Score más altos.



Modelo de predicción

Features

- Claims_Frequency
- Total_Discounts
- Inquiries
- Time_to_Conversion
- Credit_Score
- Policy_Type

Algoritmos empleados

Random Forest

Se desarrollan múltiples árboles de decisión en paralelo. El resultado es el mejor de ellos.

XGBoost

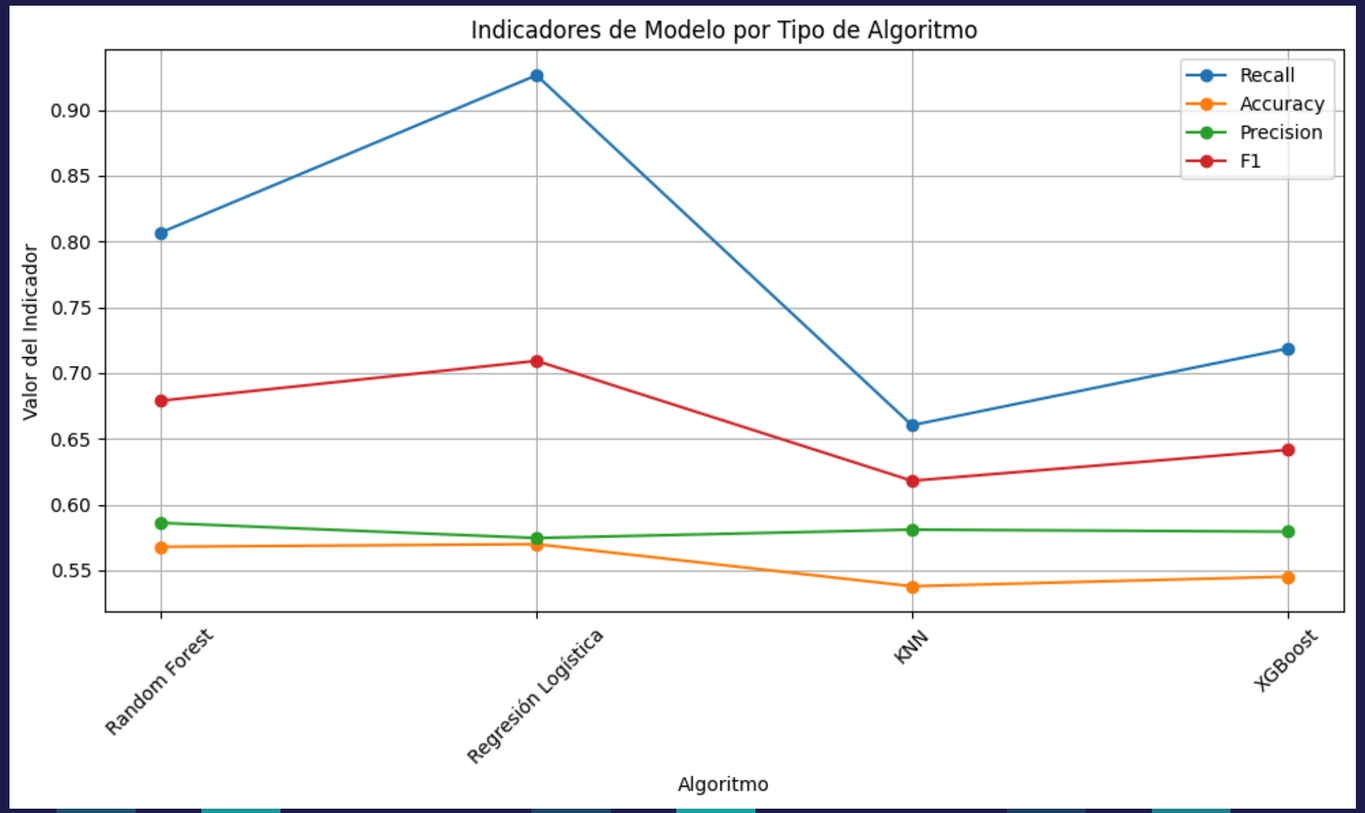
Es un algoritmo de Boosting, es decir, que está compuesto de múltiples algoritmos sencillos ejecutados en serie. Estos aprenden de los errores del anterior.

KNN

Detecta patrones y clasifica en base a los vecinos más cercanos.

Logistic Regression

Se procede con una regresión loística binaria (0; 1)



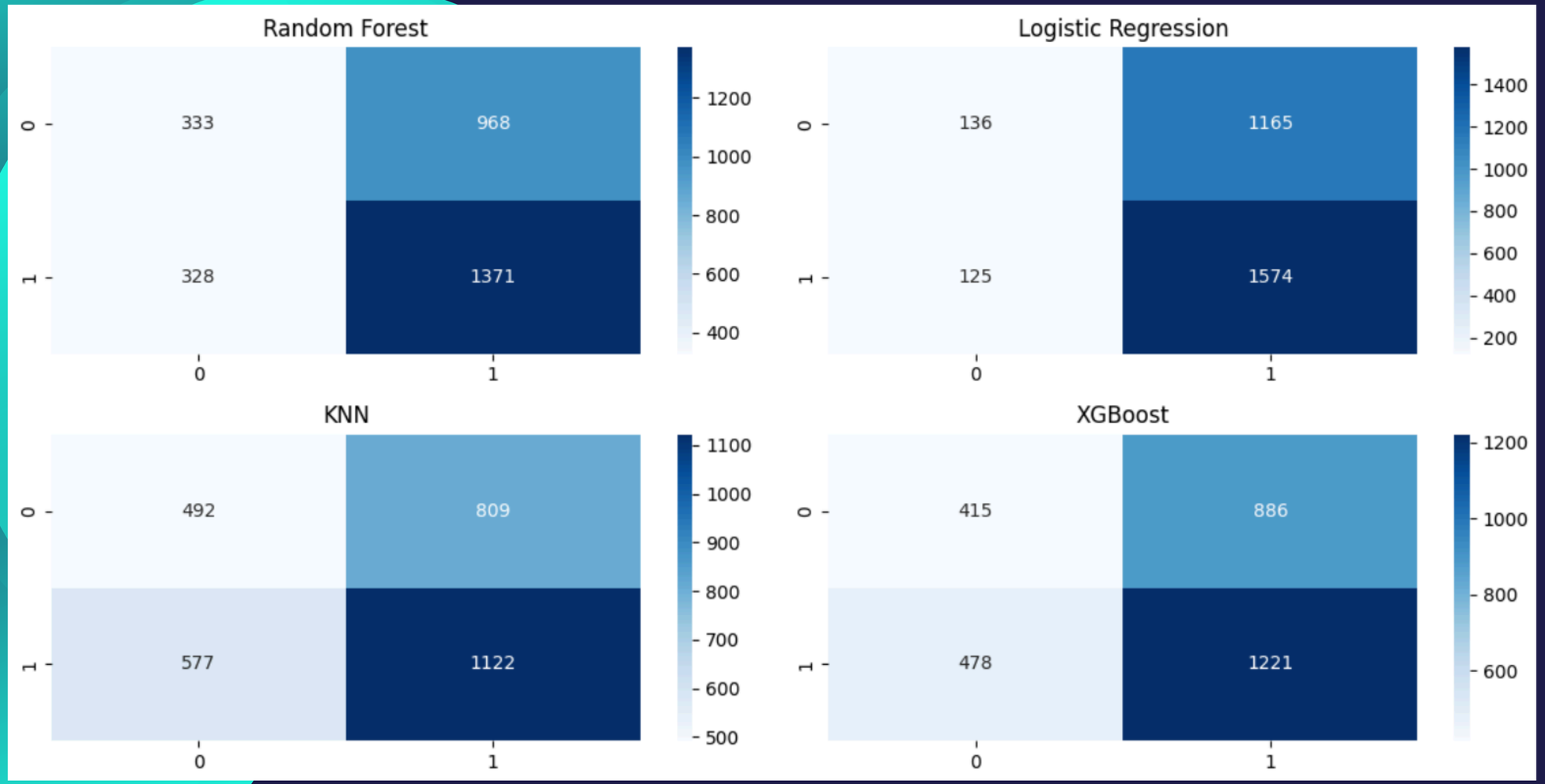
KPIs

Se destaca el alto recall de Random Forest y de Regresión Logística.

Por otro lado, KNN presenta se ve superado en todas las métricas.

Finalmente, XGBoost no tiene el Recall más alto, no obstante encuentra todas las métricas razonablemente estables.

Matriz de Confusión



Selección de modelo

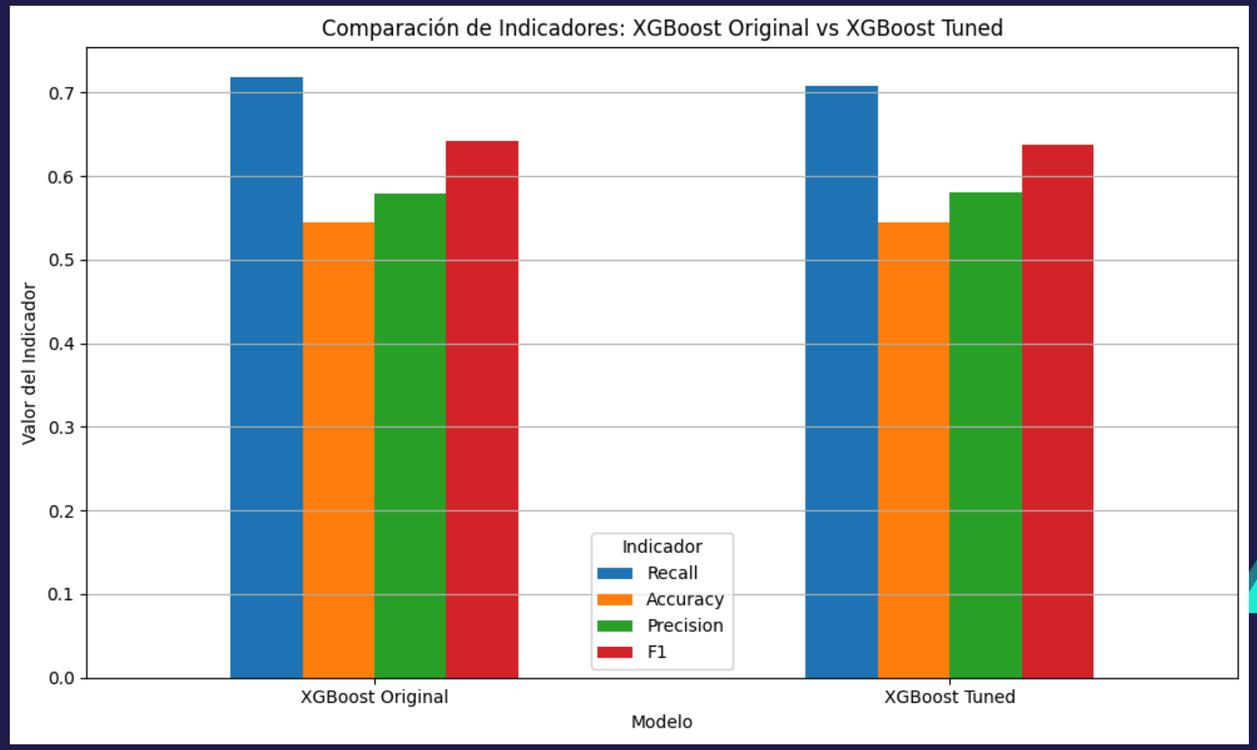
La definición de cual es el algoritmo más apropiado no es claro. Esto se debe a que el objetivo declarado al comienzo está en mejorar el profit al identificar clientes compradores y ahorrar al no considerar los no compradores.

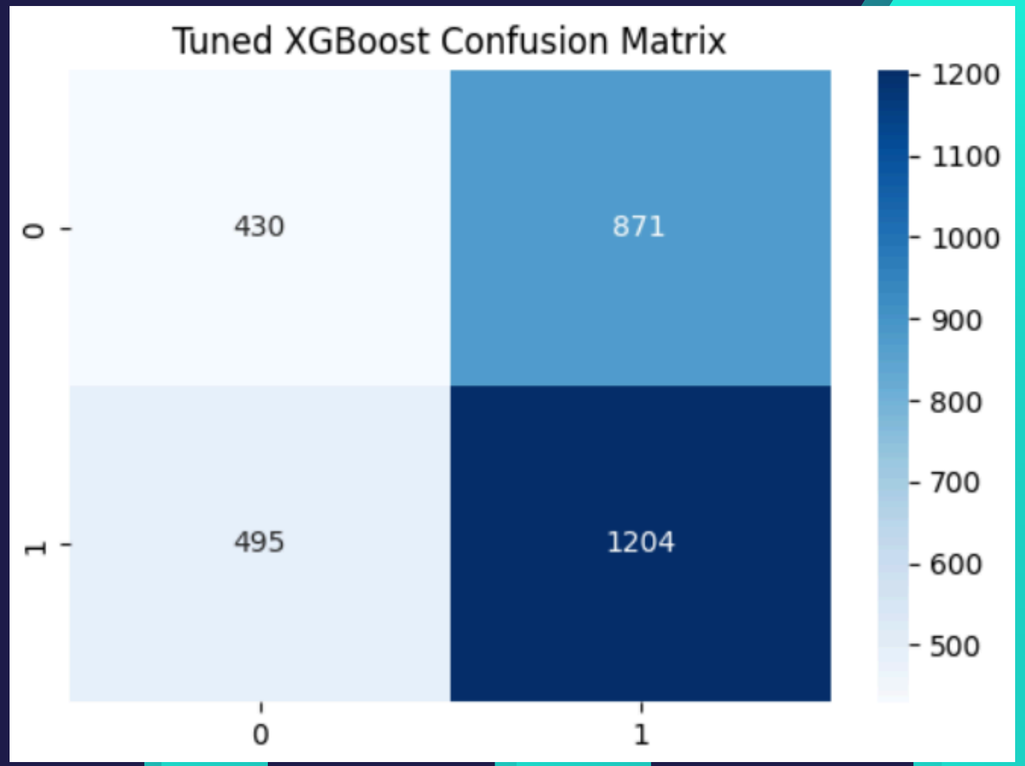
Los algoritmos cumplen la misma consigna tal que a mayor captación de clientes compradores (True Positives) también se encuentra menos ahorro en no inversión (True Negatives y False Negatives).

MODELO SELECCIONADO

Debido a sus indicadores estables, se propone continuar con XGBoost

Hiperparámetros





Conclusión

XGBoost con hiperparámetros

- Accuracy: 54%
- Recall: 71%
- Precision: 58%
- F1-Score: 64%

Se aclara que se puede optar por utilizar los otros modelos en caso de desear una postura más selectiva en clientes o más holgada. Se concluye indicando el orden de mayor a menor en exigencia de selección de clientes:

1. KNN
2. XGBoost
3. Random Forest
4. Logistic Regression

Próximos pasos

Para enriquecer el análisis del modelo, podrían utilizarse otros datos en el futuro.

Variables temporales

- En qué etapa del año se comunicó el cliente?
- A qué hora?

Variables del cliente

- En los casos que el cliente proviene de una agencia competidora, cuál es?

Variables del contacto

- Qué oficina de ventas fue la que se comunicó con el cliente?

MAURO SALINAS

Gracias

mauroasalinas@hotmail.com