Lead Scoring por medio de una Regresión Logística

Verónica Rodríguez|29 de Julio 2023

Descripción del Proyecto

Utilicé un data set que encontré en Kaggle, donde la finalidad es crear un modelo de regresión logística donde podamos puntuar a los clientes potenciales entre 0 y 100.
La más alta quiere decir que está activo y es más probable que se convierta, una más baja, determinará un cold lead y una alta probabilidad de que no se convierta.

Introducción

El negocio en cuestión es una escuela que vende diferentes cursos. Esta empresa da a conocer sus cursos en distintos websites, por medio de la herramienta Search de buscadores como Google, despúes de trasladan a sus landing pages en donde encontrarán un formulario de registro, para obtener información más detallada.

Tenemos un conjunto de datos de alrededor de 9,000 líneas. La información que se puede encontrar es cuál es la fuente del cliente, cuánto tiempo estuvo en el website, cuántas visitas, cuándo fue su última actividad, la especialización que tienen, última actividad y más.

Tenemos ya una variable de destino, en este caso, es la columna "Convertido", que indica si un cliente potencial anterior se convirtió o no, donde 1 significa que se convirtió y 0 significa que no se convirtió.

Introducción

Una vez que se adquieren estos clientes potenciales, los empleados del equipo de ventas comienzan el proceso de ventas, aquí es cuando los leads se convierten en clientes o no, sin embargo actualmente la taza de conversión de esta escuela es de alrededor del 30%. Es decir, que son más los leads que no se convierten.

Es por eso que la empresa desea identificar los clientes potenciales más potenciales, también conocidos como "Hot Leads".

Cuando ventas tenga conocimiento de esta información, se concentrará más en los leads que tienen mayor probabilidad de comprar y no dará el mismo seguimiento a todos.

Metodología

```
In [132…  data.shape

Out[132…  (9240, 37)

In [110…  data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                         Non-Null Count  Dtype  
---  ------                                         --------------  -----  
 0   Prospect ID                                    9240 non-null   object 
 1   Lead Number                                    9240 non-null   int64  
 2   Lead Origin                                    9240 non-null   object 
 3   Lead Source                                    9204 non-null   object 
 4   Do Not Email                                   9240 non-null   object 
 5   Do Not Call                                    9240 non-null   object 
 6   Converted                                      9240 non-null   int64  
 7   TotalVisits                                    9103 non-null   float64
 8   Total Time Spent on Website                    9240 non-null   int64  
 9   Page Views Per Visit                           9103 non-null   float64
 10  Last Activity                                  9137 non-null   object 
 11  Country                                        6779 non-null   object 
 12  Specialization                                 7802 non-null   object 
 13  How did you hear about X Education             7033 non-null   object 
 14  What is your current occupation                6550 non-null   object 
 15  What matters most to you in choosing a course  6531 non-null   object 
 16  Search                                         9240 non-null   object 
 17  Magazine                                       9240 non-null   object 
 18  Newspaper Article                              9240 non-null   object 
 19  X Education Forums                             9240 non-null   object 
 20  Newspaper                                      9240 non-null   object 
 21  Digital Advertisement                          9240 non-null   object 
 22  Through Recommendations                        9240 non-null   object 
 23  Receive More Updates About Our Courses         9240 non-null   object 
 24  Tags                                           5887 non-null   object 
 25  Lead Quality                                   4473 non-null   object 
 26  Update me on Supply Chain Content              9240 non-null   object 
 27  Get updates on DM Content                      9240 non-null   object 
 28  Lead Profile                                   6531 non-null   object 
 29  City                                           7820 non-null   object 
 30  Asymmetrique Activity Index                    5022 non-null   object 
 31  Asymmetrique Profile Index                     5022 non-null   object 
 32  Asymmetrique Activity Score                    5022 non-null   float64
 33  Asymmetrique Profile Score                     5022 non-null   float64
 34  I agree to pay the amount through cheque       9240 non-null   object 
 35  A free copy of Mastering The Interview         9240 non-null   object 
 36  Last Notable Activity                          9240 non-null   object 
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
```

Comenzamos con esta cantidad de líneas y columnas, sin embargo durante el proceso fuimos limpiando algunas columnas, eliminando y haciendo transformaciones en los datos.

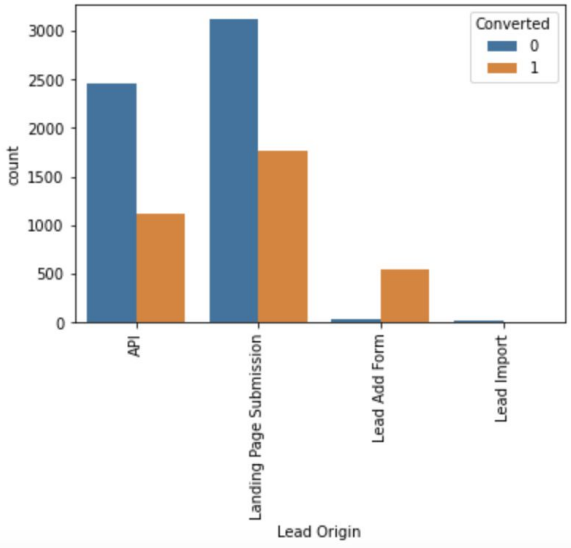
Metodología

Aquí podemos ver el Origen de los leads y en la otra columna la fuente.

Lead Origin

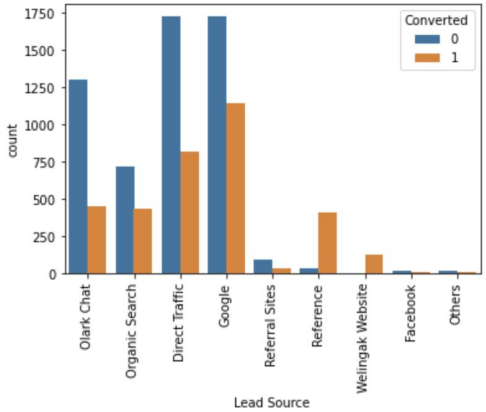
In [42]:
```
sns.countplot(x = "Lead Origin", hue = "Converted", data = data)
xticks(rotation = 90)
```

Out[42]:
```
(array([0, 1, 2, 3]),
 [Text(0, 0, 'API'),
  Text(1, 0, 'Landing Page Submission'),
  Text(2, 0, 'Lead Add Form'),
  Text(3, 0, 'Lead Import')])
```



In [45]:
```
sns.countplot(x = "Lead Source", hue = "Converted", data = data)
xticks(rotation = 90)
```

Out[45]:
```
(array([0, 1, 2, 3, 4, 5, 6, 7, 8]),
 [Text(0, 0, 'Olark Chat'),
  Text(1, 0, 'Organic Search'),
  Text(2, 0, 'Direct Traffic'),
  Text(3, 0, 'Google'),
  Text(4, 0, 'Referral Sites'),
  Text(5, 0, 'Reference'),
  Text(6, 0, 'Welingak Website'),
  Text(7, 0, 'Facebook'),
  Text(8, 0, 'Others')])
```



Metodología

Tenemos las primeras inferencias:

Inferencia de Origen
La API y el envío de páginas de destino tienen una tasa de conversión del 30-35%, pero el número de clientes potenciales que se originan a partir de ellas es considerable.
El formulario para agregar clientes potenciales tiene una tasa de conversión de más del 90%, pero el recuento de clientes potenciales no es muy alto.

Para mejorar la tasa general de conversión de clientes potenciales, debemos centrarnos más en mejorar la conversión de clientes potenciales de la API y el origen del envío de la página de destino y generar más clientes potenciales desde el formulario.

Inferencia de Fuente
Google y el tráfico directo generan el máximo número de clientes potenciales.
La tasa de conversión de clientes potenciales de referencia y clientes potenciales a través del sitio web de welingak es alta. Para mejorar la tasa general de conversión de clientes potenciales, la atención debe centrarse en mejorar la conversión de clientes potenciales del chat de Olark, la búsqueda orgánica, el tráfico directo y los clientes potenciales de Google y generar más clientes potenciales a partir del sitio web de referencia y welingak.

Metodología

Después de hacer la limpieza de los datos nos quedamos con 16 columnas para poder comenzar con el modelo.

```
data.head()
```

	Prospect ID	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Specialization	What is your current occupation	Tags	Lead Quality	City	Last Notable Activity
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	API	Olark Chat	0	0	0	0.0	0	0.0	Page Visited on Website	Others	Unemployed	Interested in other courses	Low in Relevance	Mumbai	Modified
1	2a272436-5132-4136-86fa-dcc88c88f482	API	Organic Search	0	0	0	5.0	674	2.5	Email Opened	Others	Unemployed	Ringing	Not Sure	Mumbai	Email Opened
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	Landing Page Submission	Direct Traffic	0	0	1	2.0	1532	2.0	Email Opened	Business Administration	Student	Will revert after reading the email	Might be	Mumbai	Email Opened
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	Landing Page Submission	Direct Traffic	0	0	0	1.0	305	1.0	Unreachable	Media and Advertising	Unemployed	Ringing	Not Sure	Mumbai	Modified
4	3256f628-e534-4826-9d63-4a8b88782852	Landing Page Submission	Google	0	0	1	2.0	1428	1.0	Converted to Lead	Others	Unemployed	Will revert after reading the email	Might be	Mumbai	Modified

Metodología

Convertí algunas variables en binarias y creé algunas características ficticias, para las variables que tienen varios niveles.

Data Preparation

Converting some binary variables (Yes/No) to 1/0

In [49]:

```
# List of variables to map

varlist =  ['Do Not Email', 'Do Not Call']

# Defining the map function
def binary_map(x):
    return x.map({'Yes': 1, "No": 0})

# Applying the function to the housing list
data[varlist] = data[varlist].apply(binary_map)
```

For categorical variables with multiple levels, create dummy features (one-hot encoded)

In [50]:

```
# Creating a dummy variable for some of the categorical variables and dropping the first one.
dummy1 = pd.get_dummies(data[['Lead Origin', 'Lead Source', 'Last Activity', 'Specialization','What is your current occupation',
                              'Tags','Lead Quality','City','Last Notable Activity']], drop_first=True)
dummy1.head()
```

Out[50]:

	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Facebook	Lead Source_Google	Lead Source_Olark Chat	Lead Source_Organic Search	Lead Source_Others	Lead Source_Reference	Lead Source_Referral Sites	Lead Source_Welingak Website	Last Activity_Converted to Lead	Ac
0	0	0	0	0	0	1	0	0	0	0	0	0	
1	0	0	0	0	0	0	1	0	0	0	0	0	
2	1	0	0	0	0	0	0	0	0	0	0	0	
3	1	0	0	0	0	0	0	0	0	0	0	0	
4	1	0	0	0	1	0	0	0	0	0	0	1	

Metodología

Aquí podemos ver que la tasa de conversión es de casi 38%

Converted

Converted is the target variable, Indicates whether a lead has been successfully converted (1) or not (0).

In [41]:
```
Converted = (sum(data['Converted'])/len(data['Converted'].index))*100
Converted
```

Out[41]:
```
37.85541106458012
```

Metodología

Estamos construyendo el modelo de Regresión

Model Building

Running Your First Training Model

In [62]:
```
import statsmodels.api as sm
```

In [63]:
```
# Logistic regression model
logm1 = sm.GLM(y_train,(sm.add_constant(X_train)), family = sm.families.Binomial())
logm1.fit().summary()
```

Out[63]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6248
Model Family:	Binomial	**Df Model:**	102
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1237.2
Date:	Fri, 25 Aug 2023	**Deviance:**	2474.4
Time:	18:06:43	**Pearson chi2:**	3.66e+04
No. Iterations:	24		
Covariance Type:	nonrobust		

Metodología

Estamos construyendo el modelo de Regresión

	coef	std err	z	P>|z|	[0.025	0.975]
const	24.3058	2.19e+05	0.000	1.000	-4.29e+05	4.29e+05
Do Not Email	-1.4782	0.334	-4.422	0.000	-2.133	-0.823
Do Not Call	23.8191	1.36e+05	0.000	1.000	-2.67e+05	2.67e+05
TotalVisits	0.1611	0.036	4.429	0.000	0.090	0.232
Total Time Spent on Website	1.1488	0.064	17.870	0.000	1.023	1.275
Page Views Per Visit	-0.1582	0.074	-2.148	0.032	-0.303	-0.014
Lead Origin_Landing Page Submission	-1.0249	0.223	-4.606	0.000	-1.461	-0.589
Lead Origin_Lead Add Form	-0.2415	1.330	-0.181	0.856	-2.849	2.366
Lead Origin_Lead Import	29.8727	2.08e+05	0.000	1.000	-4.07e+05	4.07e+05
Lead Source_Facebook	-28.6509	2.08e+05	-0.000	1.000	-4.07e+05	4.07e+05
Lead Source_Google	0.1843	0.155	1.186	0.236	-0.120	0.489
Lead Source_Olark Chat	0.9945	0.228	4.365	0.000	0.548	1.441
Lead Source_Organic Search	0.1513	0.210	0.720	0.471	-0.261	0.563
Lead Source_Others	0.8169	0.812	1.006	0.314	-0.775	2.408
Lead Source_Reference	1.8341	1.391	1.319	0.187	-0.892	4.560
Lead Source_Referral Sites	-0.1439	0.490	-0.293	0.769	-1.105	0.817
Lead Source_Welingak Website	5.4505	1.526	3.571	0.000	2.459	8.442
Last Activity_Converted to Lead	-19.1909	9.63e+04	-0.000	1.000	-1.89e+05	1.89e+05
Last Activity_Email Bounced	-19.5269	9.63e+04	-0.000	1.000	-1.89e+05	1.89e+05
Last Activity_Email Link Clicked	-18.3726	9.63e+04	-0.000	1.000	-1.89e+05	1.89e+05
Last Activity_Email Marked Spam	0.3186	1.28e+05	2.49e-06	1.000	-2.5e+05	2.5e+05
Last Activity_Email Opened	-19.2023	9.63e+04	-0.000	1.000	-1.89e+05	1.89e+05
Last Activity_Email Received	3.1686	2.62e+05	1.21e-05	1.000	-5.13e+05	5.13e+05
Last Activity_Form Submitted on Website	-19.0374	9.63e+04	-0.000	1.000	-1.89e+05	1.89e+05
Last Activity_Had a Phone Conversation	-16.3250	9.63e+04	-0.000	1.000	-1.89e+05	1.89e+05

Metodología

Este código utiliza la biblioteca Scikit-Learn para realizar la selección de características utilizando el método Recursive Feature Elimination (RFE) en un modelo de regresión logística.

Feature Selection Using RFE

In [64]:
```
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()

from sklearn.feature_selection import RFE
rfe = RFE(logreg, 15)             # running RFE with 15 variables as output
rfe = rfe.fit(X_train, y_train)
```

In [65]:
```
rfe.support_
```

Out[65]:
```
array([ True, False, False, False, False, False,  True, False, False,
       False, False, False, False, False, False,  True, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False,  True,  True, False, False, False, False, False, False,
       False,  True, False, False, False,  True, False, False, False,
       False,  True, False,  True,  True, False,  True,  True, False,
       False,  True,  True, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False,  True, False, False, False])
```

Metodología

Esto proporciona una visión clara de qué características fueron seleccionadas, cuáles no y cómo se clasificaron en términos de importancia relativa.

In [66]:
```
list(zip(X_train.columns, rfe.support_, rfe.ranking_))
```

Out[66]:
```
[('Do Not Email', True, 1),
 ('Do Not Call', False, 44),
 ('TotalVisits', False, 55),
 ('Total Time Spent on Website', False, 4),
 ('Page Views Per Visit', False, 57),
 ('Lead Origin_Landing Page Submission', False, 21),
 ('Lead Origin_Lead Add Form', True, 1),
 ('Lead Origin_Lead Import', False, 3),
 ('Lead Source_Facebook', False, 56),
 ('Lead Source_Google', False, 60),
 ('Lead Source_Olark Chat', False, 6),
 ('Lead Source_Organic Search', False, 61),
 ('Lead Source_Others', False, 58),
 ('Lead Source_Reference', False, 82),
 ('Lead Source_Referral Sites', False, 45),
 ('Lead Source_Welingak Website', True, 1),
 ('Last Activity_Converted to Lead', False, 43),
 ('Last Activity_Email Bounced', False, 39),
 ('Last Activity_Email Link Clicked', False, 81),
 ('Last Activity_Email Marked Spam', False, 53),
 ('Last Activity_Email Opened', False, 47),
 ('Last Activity_Email Received', False, 80),
 ('Last Activity_Form Submitted on Website', False, 51),
 ('Last Activity_Had a Phone Conversation', False, 16),
 ('Last Activity_Olark Chat Conversation', False, 18),
 ('Last Activity_Page Visited on Website', False, 89),
 ('Last Activity_Resubscribed to emails', False, 77),
 ('Last Activity_SMS Sent', False, 8),
 ('Last Activity_Unreachable', False, 19),
 ('Last Activity_Unsubscribed', False, 24),
 ('Last Activity_View in browser link Clicked', False, 40),
 ('Last Activity_Visited Booth in Tradeshow', False, 90),
 ('Specialization_Business Administration', False, 86),
 ('Specialization_E-Business', False, 85),
 ('Specialization_E-COMMERCE', False, 20),
 ('Specialization_Finance Management', False, 50),
 ('Specialization_Healthcare Management', False, 49),
 ('Specialization_Hospitality Management', False, 83),
 ('Specialization_Human Resource Management', False, 59),
```

Metodología

Aquí discrimino ya solo las características seleccionadas como imporantes.

In [67]:

```
col = X_train.columns[rfe.support_]
col
```

Out[67]:

```
Index(['Do Not Email', 'Lead Origin_Lead Add Form',
       'Lead Source_Welingak Website', 'Tags_Busy', 'Tags_Closed by Horizzon',
       'Tags_Lost to EINS', 'Tags_Ringing',
       'Tags_Will revert after reading the email', 'Tags_invalid number',
       'Tags_number not provided', 'Tags_switched off',
       'Tags_wrong number given', 'Lead Quality_Not Sure',
       'Lead Quality_Worst', 'Last Notable Activity_SMS Sent'],
      dtype='object')
```

Metodología

El resultado del resumen me dio información detallada sobre cómo cada característica está relacionada con la variable objetivo y cómo el modelo se ajusta a los datos de entrenamiento.

dtype= object)

In [69]:

```
X_train_sm = sm.add_constant(X_train[col])
logm2 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm2.fit()
res.summary()
```

Out[69]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6335
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1588.2
Date:	Fri, 25 Aug 2023	**Deviance:**	3176.4
Time:	18:11:12	**Pearson chi2:**	3.57e+04
No. Iterations:	24		
Covariance Type:	nonrobust		

Metodología

- Los positivos: "Lead Origin_Lead Add Form", "Lead Source_Welingak Website", "Tags_Busy", "Tags_Closed by Horizzon", "Tags_Lost to EINS", "Tags_Will revert after reading the email", "Last Notable Activity_SMS Sent" sugieren un impacto positivo en la probabilidad de conversión.
- Los negativos como "Do Not Email", "Tags_Ringing", "Tags_switched off", "Lead Quality_Not Sure", "Lead Quality_Worst" sugieren que tienen un impacto negativo en la probabilidad de conversión.
- Los intervalos de confianza al 95% proporcionan un rango en el que se espera que caiga el valor real del coeficiente con un 95% de confianza.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5851	0.207	-7.673	0.000	-1.990	-1.180
Do Not Email	-1.2714	0.214	-5.946	0.000	-1.690	-0.852
Lead Origin_Lead Add Form	1.1288	0.363	3.110	0.002	0.417	1.840
Lead Source_Welingak Website	3.3635	0.818	4.112	0.000	1.760	4.967
Tags_Busy	3.4323	0.332	10.345	0.000	2.782	4.083
Tags_Closed by Horizzon	7.6671	0.761	10.076	0.000	6.176	9.158
Tags_Lost to EINS	8.8786	0.752	11.807	0.000	7.405	10.352
Tags_Ringing	-2.2142	0.338	-6.543	0.000	-2.877	-1.551
Tags_Will revert after reading the email	3.6531	0.225	16.267	0.000	3.213	4.093
Tags_invalid number	-23.7322	2.2e+04	-0.001	0.999	-4.31e+04	4.31e+04
Tags_number not provided	-24.2210	3.81e+04	-0.001	0.999	-7.48e+04	7.47e+04
Tags_switched off	-2.8592	0.588	-4.860	0.000	-4.012	-1.706
Tags_wrong number given	-23.3279	3.14e+04	-0.001	0.999	-6.16e+04	6.15e+04
Lead Quality_Not Sure	-3.5346	0.126	-28.071	0.000	-3.781	-3.288
Lead Quality_Worst	-3.9822	0.846	-4.709	0.000	-5.640	-2.325
Last Notable Activity_SMS Sent	2.8226	0.123	22.908	0.000	2.581	3.064

Metodología

In [77]:

```
# Getting the predicted values on the train set
y_train_pred = res.predict(X_train_sm)
y_train_pred[:10]
```

Out[77]:

```
3009    0.189083
1012    0.061798
9226    0.000691
4750    0.786646
7987    0.977215
1281    0.992107
2880    0.189083
4971    0.752206
7536    0.888252
1248    0.000691
dtype: float64
```

Explicación

Distribución de las Observaciones y Modelo de Regresión:

- El modelo se ajustó a un total de 6,336 observaciones.
- Se utilizó el método IRLS (Iteratively Reweighted Least Squares) para ajustar el modelo.
- El modelo se basa en la función de enlace logit.

Coeficientes de Regresión:

- Cada coeficiente de regresión se asocia a una variable independiente en el modelo.
- Los coeficientes muestran cómo un cambio en una variable independiente afecta la probabilidad log-odds de que la variable de respuesta sea 1 (Converted).

Significancia Estadística:

- Cada coeficiente tiene su p-value asociado.
- Los p-values indican si un coeficiente es estadísticamente significativo para predecir la variable de respuesta.
- Un p-value pequeño (generalmente < 0.05) indica que el coeficiente es significativo.

Impacto y Dirección:

- Coeficientes positivos indican que un incremento en la variable independiente aumenta la log-odds de que la respuesta sea 1.
- Coeficientes negativos indican que un incremento en la variable independiente disminuye la log-odds de que la respuesta sea 1.

Variables Importantes:

- Las variables con coeficientes significativos tienen un impacto en la predicción de la variable de respuesta.
- Las variables más importantes en este modelo incluyen "Tags_Closed by Horizzon", "Tags_Lost to EINS",

Explicación

Distribución de las Observaciones y Modelo de Regresión:

- El modelo se ajustó a un total de 6,336 observaciones.
- Se utilizó el método IRLS (Iteratively Reweighted Least Squares) para ajustar el modelo.
- El modelo se basa en la función de enlace logit.

Coeficientes de Regresión:

- Cada coeficiente de regresión se asocia a una variable independiente en el modelo.
- Los coeficientes muestran cómo un cambio en una variable independiente afecta la probabilidad log-odds de que la variable de respuesta sea 1 (Converted).

Significancia Estadística:

- Cada coeficiente tiene su p-value asociado.
- Los p-values indican si un coeficiente es estadísticamente significativo para predecir la variable de respuesta.
- Un p-value pequeño (generalmente < 0.05) indica que el coeficiente es significativo.

Impacto y Dirección:

- Coeficientes positivos indican que un incremento en la variable independiente aumenta la log-odds de que la respuesta sea 1.
- Coeficientes negativos indican que un incremento en la variable independiente disminuye la log-odds de que la respuesta sea 1.

Variables Importantes:

- Las variables con coeficientes significativos tienen un impacto en la predicción de la variable de respuesta.
- Las variables más importantes en este modelo incluyen "Tags_Closed by Horizzon", "Tags_Lost to EINS", "Tags_Will revert after reading the email", entre otras.

Importancia Relativa de Variables:

- Al eliminar algunas variables con p-values altos (no significativas), el modelo se simplifica sin una pérdida significativa de rendimiento.
- Esto puede resultar en un modelo más interpretable y fácil de usar.

Desempeño del Modelo:

- La medida de ajuste log-likelihood indica la bondad de ajuste del modelo. A mayor valor, mejor ajuste.
- El deviance muestra cuánto el modelo difiere de un modelo ideal. Menor deviance es mejor ajuste.

Conclusiones

Conclusiones

Estos valores son el resultado de aplicar el modelo de regresión logística a las características (variables independientes) de los registros en el conjunto de entrenamiento y calcular las probabilidades log-odds resultantes. En muchas aplicaciones, estos valores se pueden convertir nuevamente a probabilidades en la escala de 0 a 1 para interpretar más intuitivamente las predicciones del modelo.

Conclusiones

El resultado final es una tabla que contiene información sobre las predicciones del modelo y los valores reales correspondientes para cada registro en el conjunto de entrenamiento. Esto es útil para realizar análisis comparativos, calcular métricas de evaluación y comprender cómo se ajustan las predicciones del modelo a los valores reales.

Conclusiones

	Converted	Converted_prob	Prospect ID
0	0	0.189083	3009
1	0	0.061798	1012
2	0	0.000691	9226
3	1	0.786646	4750
4	1	0.977215	7987

Las filas 0, 1 y 2 tienen predicciones de "No conversión" con probabilidades bajas. Esto significa que el modelo está bastante seguro de que estos registros no se convertirán.

Las filas 3 y 4 tienen predicciones de "Conversión" con probabilidades relativamente altas. Esto indica que el modelo está bastante seguro de que estos registros se convertirán.

Conclusiones

Usé la matriz de confusión para evaluar el rendimiento del modelo de clasificación y comprender cómo está acertando o fallando en la predicción de diferentes clases.

In [81]:
```
from sklearn import metrics

# Confusion matrix
confusion = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.predicted )
print(confusion)
```

```
[[3753  152]
 [ 357 2089]]
```

In [82]:
```
# Let's check the overall accuracy.
print(metrics.accuracy_score(y_train_pred_final.Converted, y_train_pred_final.predicted))
```

```
0.9198551409226894
```

El resultado 0.9198551409226894 es el valor de la precisión calculado para el modelo de predicción. Este valor representa la proporción de predicciones correctas realizadas por el modelo en relación con el total de predicciones.
En este contexto, una precisión del 0.92 (o alrededor del 92%) significa que aproximadamente el 92% de las predicciones realizadas por el modelo coinciden con las etiquetas reales en el conjunto de datos.

Esto indica que el modelo tiene un buen rendimiento en términos de la precisión general de sus predicciones.

Conclusiones

La mayoría de los valores de VIF son relativamente bajos, lo que sugiere que la multicolinealidad entre estas variables, no es un problema significativo en el modelo.

Out[84]:

	Features	VIF
12	Last Notable Activity_SMS Sent	2.71
7	Tags_Will revert after reading the email	2.66
1	Lead Origin_Lead Add Form	1.58
6	Tags_Ringing	1.55
2	Lead Source_Welingak Website	1.34
4	Tags_Closed by Horizzon	1.13
0	Do Not Email	1.12
3	Tags_Busy	1.11
10	Lead Quality_Not Sure	1.10
5	Tags_Lost to EINS	1.04
8	Tags_number not provided	1.04
11	Lead Quality_Worst	1.02
9	Tags_switched off	1.01

Conclusiones

En el contexto de evaluación de modelos de clasificación, el valor "0.9321731369924141" que has proporcionado representa la tasa de verdaderos positivos (True Positive Rate), también conocida como sensibilidad o recall.

```
# positive predictive value
print (TP / float(TP+FP))
```

```
0.9321731369924141
```

Recomendaciones Finales

- Realizar llamadas a los clientes potenciales provenientes de las fuentes de clientes potenciales "Sitios web de Welingak" y "Referencia", ya que es más probable que se conviertan.
- Llamar a los clientes potenciales que son los "working professionals", ya que es más probable que se conviertan.
- Contactar a los clientes potenciales que pasaron "más tiempo en los sitios web", ya que es más probable que se conviertan.
- Llamar a los clientes potenciales provenientes de las fuentes de clientes potenciales "Olark Chat", ya que es más probable que se conviertan.
- La empresa debe llamar a los clientes potenciales cuya última actividad fue el envío de SMS, ya que es más probable que se conviertan.
- La empresa no debe brindar toda su atención los clientes potenciales cuyo origen sea "Envío de página de destino", ya que es poco probable que se conviertan.
- No se recomienda llamar a los clientes potenciales cuya especialización era "Otros", ya que es poco probable que se conviertan.
- La empresa no debe realizar llamadas a quienes eligieron la opción "No enviar correo electrónico" como "sí", es poco probable que se conviertan.